National Assessment Governing Board

Committee on Standards, Design and Methodology

March 2-3, 2017

AGENDA

Thursday, March 2		
1:30 – 1:35 pm	Welcome, Introductions, and Agenda Overview *Andrew Ho, Chair*	
1:35 – 1:50 pm	2017 Writing Grade 4 Achievement Levels Setting Project Update *Timothy O'Neil, Pearson*	Attachment A
1:50 – 2:30 pm	Uses of NAEP *Ina Mullis, Boston College*	Attachment B
2:30 – 4:00 pm	Follow up to Evaluation of NAEP Achievement Levels • Governing Board Response • Planned Procurements • Revising the Governing Board Policy on Developing Student Performance Levels on NAEP *Andrew Ho, Chair* *Sharyn Rosenberg, Assistant Director for Psychometrics*	Attachment C
	Information Item • Update on Transition to Digital Based Assessments	Attachment D

National Assessment Governing Board

Committee on Standards, Design and Methodology

March 2-3, 2017

AGENDA

Friday, March 3		
10:00 – 11:00 am	Joint Session with the Assessment Development Committee (ADC) on Dynamic Frameworks *Andrew Ho, COSDAM Chair* *Cary Sneider, ADC Vice Chair*	Attachment E
11:00 am – 12:00 pm	COSDAM Role in Implementing the Strategic Vision *Andrew Ho, Chair*	See materials sent under separate cover

Developing Achievement Levels for the National Assessment of Educational Progress Writing at Grade 4

Purpose: The purpose of this session is to provide an update to the Committee on Standards, Design and Methodology (COSDAM) regarding the development of achievement levels for the 2017 NAEP Grade 4 Writing. In this session, Tim O'Neil, NAEP Grade 4 Writing Achievement Levels-Setting (ALS) Project Director for Pearson, will provide a brief update on the project.

Legend:
Light shading: Completed
Dark shading: Current status
No shading: To be completed after 3/01/2017

Purpose: The purpose of this session is to provide an update to the Committee on Standards, Design and Methodology (COSDAM) regarding the development of achievement levels for 2017 NAEP Grade 4 Writing. In this session, Dr. Tim O'Neil, NAEP Grade 4 Writing Achievement Levels-Setting (ALS) Project Director for Pearson, will provide a brief update on the project.

Project Overview: On August 3, 2016, the National Assessment Governing Board (Governing Board) awarded a contract to Pearson (as a result of a competitive bidding process) for developing achievement levels for the National Assessment of Educational Progress (NAEP) for grade 4 writing. The 2017 Grade 4 NAEP Writing assessment is the first administration of the grade 4 assessment developed to meet the design specifications described in the current computer-based Writing Framework. The assessment is a digital-based assessment, comprised of constructed response items, for which students compose and construct their responses using word processing software on a tablet. The assessment is to be administered to a nationally representative sample of approximately 22,000 grade 4 students in the spring of 2017.[1]

Dr. Tim O'Neil is the grade 4 writing ALS project director at Pearson and Dr. Marc Johnson is the assistant project director at Pearson. Pearson will conduct a field trial, a pilot study, and an achievement levels-setting (ALS) meeting and produce a set of recommendations for the Governing Board to consider in establishing achievement levels for the grade 4 NAEP writing assessment. The Governing Board is expected to take action on the writing grade 4 achievement levels during the May 2018 meeting. Pearson will utilize a body of work methodology using Moodle software to collect panelist ratings and present feedback. Dr. Lori Nebelsick-Gullet will serve as the process facilitator for the pilot and operational ALS meetings; Victoria Young will serve as the content facilitator for the pilot and operational ALS meetings; and Drs. Susan Cooper Loomis and Steven Fitzpatrick will serve as consultants.

For setting standards, Pearson will use a body of work approach in which panelists will make content-based cut score recommendations. The body of work methodology is a holistic standard setting method for which panelists evaluate sets of examinee work (i.e., bodies of work) and provide a holistic judgment about each student set. These content-based judgments will be made over three rounds. The process to be implemented for the standard setting meeting follows body of work procedures used in previous NAEP standard setting studies. In addition, a field trial will be conducted prior to the pilot study which will provide an opportunity to try out a number of key aspects of the ALS plan, including the logistical design of the ALS studies such as the use of tablets and laptop computers, the ease with which the panelists can enter judgments and questionnaire responses, and the arrangement of tables and panelists.

The Governing Board policy on Developing Student Performance Levels for NAEP (https://www.nagb.org/content/nagb/assets/documents/policies/developing-student-performance.pdf) requires appointment of a committee of technical advisors who have expertise in standard setting and psychometrics in general, as well as issues specific to NAEP. These advisors will be convened for 8 in-person meetings and up to 6 webinars to provide advice at every key point in the process. They provide feedback on plans and materials before activities

[1] Achievement levels were set for Writing grades 8 and 12 with the 2011 administration of those assessments. The grade 4 assessment initially was scheduled to be administered in 2013 but the Governing Board postponed it to 2017 due to budgetary constraints.

are implemented and review results of the process and analyses. Six external experts in standard setting are serving on the Technical Advisory Committee on Standard Setting (TACSS):

Dr. Gregory Cizek
Professor of Educational Measurement, University of North Carolina at Chapel Hill

Dr. Barbara Dodd
Professor of Professor of Quantitative Methods, University of Texas at Austin

Dr. Steve Ferrara
Independent Consultant

Dr. Matthew Johnson
Associate Professor of Statistics and Education, Teachers College, Columbia University

Dr. Vaughn G. Rhudy
Executive Director, Office of Assessment, West Virginia Department of Education

Dr. Mary Pitoniak
Senior Strategic Advisor for Statistical Analysis, Data Analysis, and Psychometric Research, Educational Testing Service (NAEP Design, Analysis, and Reporting Contractor)

March 2017 Update:

Public Comment on the Design Document

The Design Document is intended to provide the foundation for all achievement levels-setting activities. The Design Document for the NAEP Grade 4 Writing achievement levels-setting process includes discussion of the methodology, procedures, and documentation of the entire project. COSDAM reviewed a draft of the Design Document at the November 2016 meeting, and the document was edited following that discussion.

The Design Document was distributed for public comment from January 10th to February 10th, 2017 at (http://downloads.pearsonassessments.com/naeptelassessment/). Since past efforts to obtain comment on ALS design documents has tended not to result in many comments, Pearson implemented a more aggressive outreach plan that focused on content experts, measurement experts, and education-related organizations known to use NAEP ALS data. In particular, individuals who made presentations about standard setting at the annual meeting of the National Council on Measurement in Education and the American Education Research Association in the past five to 10 years were contacted for input. The latter appears to have been an effective strategy, as several comments were received from this stakeholder group.

Comments were reviewed and discussed with the TACSS during the February 16th webinar. Pearson is making edits to the Design Document to address some of the comments received.

Update on Preparations for the Field Trial

Pearson is currently in the process of creating all materials and tools necessary to conduct the field trial, to include creation of the Moodle interface. Completed materials and the Moodle interface were reviewed and discussed with the TACSS during the February 16th webinar.

The field trial will be conducted from June 5-6, 2017. The location was changed from Pearson's Austin office to their office in San Antonio due to space availability constraints. This change is not expected to negatively impact recruitment (where panelists will be targeted within a 30-50 mile radius).

During the March 2017 COSDAM session, Writing ALS Project Director Tim O'Neil will provide an update on preparations for the June 2017 field trial meeting as well as a summary of public comments received and recommendations from the February TACSS webinar.

Public Comment 1

I have read the standard setting plan for Grade 4 Writing. The document is very detailed and clear. However, I read also documents by Measured Progress (Measured Progress, 2012a, 2012b) that were mentioned in the Design Document to supplement my background. My comment is as follows:

The challenge of setting actual cut scores in a NAEP context is to identify the score levels on the NAEP score reporting metric that demarcate the different proficiency levels that have been previously defined at a conceptual level. In this case, the assessment consists of two writing performance tasks. Given that each task is scored by a six-point rubric, in principle, there are 36 possible score patterns for each possible pairing of writing tasks. The EAP is computed to summarize the performance on the two tasks into a single number. (I'm assuming the process has access to the necessary NAEP data such that the EAP is based on the response pattern *and* the background information used in group-level analyses.) From there, a set of BoWs spanning the EAP score range are given to the panelists to sort into the predefined achievement levels. The BoW method, as implemented here, method calls for a range finding and pinpointing phase based *solely* on the EAP associated with the BoWs.

Although the approach seems reasonable to me, there is a strong reliance on the adequate fit of the IRT model to these data. In addition, it appears that the score on each task is the result of human scoring. While it may be necessary to assume a reasonable good fit of the IRT model and that the human scoring is sound, it may also be possible to not fully rely on those assumptions. Specifically, during the range finding and pinpointing phases, rather than simply rely on the EAP as a way to index students, it could be useful to explicitly include atypical BoWs where the performance on the two tasks differ markedly, although, interestingly, in page 30 it is stated that such responses will be explicitly excluded and not presented to the panelists.

That is, the student performance can be located on a matrix like this, which could be useful for locating less typical BoWs:

	Task 2					
	Score 1	Score 2	Score 3	Score 4	Score 5	Score 6
Task 1: Score 1						
Task 1: Score 2						
Task 1: Score 3						
Task 1: Score 4						
Task 1: Score 5						
Task 1: Score 6						

To the extent a unidimensional models fits, cases with very divergent performance on the two tasks would be rare but could be more prevalent than expected since the assessment is scored by humans. From experience with human scorers, we know that raters can, on occasion, assign quite different scores to the same performance. The reason is that the scoring of writing is very cognitively demanding and there are rater tendencies, such as leniency, that exacerbate the problem. (I'm assuming that the two tasks a given student responded to are scored by different

raters.) A partial solution would be to score the performance by means of the equivalent of the hierarchical rater model (HRM, Patz, Junker, Johnson, & Mariano, 2002), which would at least remove rater effects. Clearly, that is not feasible at the moment but precisely for that reason I'm thinking that the selection of BoWs should not be guided exclusively by the EAP scores but could systematically include less typical BoWs. Alternatively, as part of the validation phase, an analysis could be performed to check on the adequacy of the tentative standards by having panelist corroborate the assignment of achievement levels to less typical performances.

In short, I appreciated the level of detail provided by the Design Document that made it possible for me to comment on the design.

Measured Progress (2012a). *Developing achievement levels on the National Assessment of Educational Progress in writing grades 8 and 12 in 2011: Process report*. Dover, NH: Author.

Measured Progress (2012b). *Developing achievement levels on the National Assessment of Educational Progress in writing grades 8 and 12: Technical report*. Dover, NH: Author.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The Hierarchical Rater Model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics in Medicine, 27*, 341-384.

Public Comment 2

Thank you for the opportunity to review the plan for standard setting for the NAEP Grade 4 Writing test. Overall, the plan looks very much like the plan Measured Progress followed in setting cut scores for grades 8 and 12 in 2011. Thus, my comments are few and apply to the earlier report as much as to your plan:

p. 15. The actual scoring procedures are a bit vague (as they were in the Measured Progress report on the 2011 standard setting). More detail is needed here. For example, what are the key criteria (3-6 perhaps) for selection into the teacher pool, and what weights should each have, and what will be the score scale for each? What research impinges on this selection process? Pearson did a nice job in recruiting and selecting panelists for the PARCC standard setting. Perhaps the same or similar procedures could be used here. At any rate, they need to be explained more fully.

p. 23. This training needs to be spelled out in some detail, either here or in a subsequent section. Again, for the record, MP did not spell out the training either.

p. 33 (Review ALDs). This process needs to be defined. (See above re MP.)

p. 35 (logistic regression). This is a deviation from the original design of BoW (cf. Kingston, et al. in Greg Cizek's 2001 or 2012 edition of Setting Performance Standards). Traditionally, Round 1 has been used to identify regions of cut scores without the use of logistic regression. It appears that this was how Measured Progress employed BoW in 2011, but the departure from the original design was not explained in that report either.

p. 36 (Round 2). It is not clear whether these are the same BoWs as in Round 1 or some new and some old to narrow the search for cut points.

p. 48. "However, mean cut scores will also be provided to panelists for complete information and in comparison to the medians." To what end?

p. 52. Therefore, there are no concrete plans for external validation at this time. Is that correct?

Public Comment 3

Page	Text/Reference	Comment
6	The use of computers strengthens the validity argument for the ALS	I think it's fine to use the computers here for logistical efficiency, but I think it's a major stretch to claim that this adds to the validity of the process
8	Pearson carefully considered the research conducted and lessons learned from previous panel studies for NAEP writing ALS projects and designed the grade 4 writing ALS process to address the fidelity of writing ALS results across grades by closely following procedures successfully implemented for grades 8 and 12 (Measured Progress, 2012a).	I don't think this is an issue of concern to you, but I strongly endorse not changing the basic methodology here from that used for the Gr. 8/12 tests.
9	Using the same approach for grade 4 writing will provide consistency across the ALS procedures implemented for the NAEP writing assessments and removes the potential for differences in the achievement level cut scores due to the use of different standard setting methods.	agree strongly
10	(http://nces.ed.gov/nationsreportcard/writing/lessons/performance.aspx). While the information provided from these sources is not likely to eliminate all concerns, it provides an empirical and factual basis for panelists to consider when making their judgments.	I guess I see no reason not to share any anecdotal/survey-type info with the panel, I'm not sure how you'd instruct panelists to "take this information into account" in doing their work. SHould I raise/lower my standards based on what a few - or even many - test-takers tell me about the experience??
11	Pearson will recruit a total of 75 panelists—20 for the field trial, 22 for the pilot study, 33 for the operational ALS meeting.	seems to me like more than plenty of panelists. I've not gone back to the earlier 8/12 reports, but assume similar sized panels were used there?
11	Panels will reflect an overall balance of gender	I assume this
11	Classroom teachers currently engaged in writing instruction at grade 4 will compose 55 percent of each panel.	sorry to be picky here, but 55% is pretty silly when the panel size is 30. Are you really thinking about half of a panelist? Why not just say that teachers will make up the plurality? Or at least add "about" prior to 55%

Page	Text/Reference	Comment
11	Representatives of non-educator groups will compose approximately 30 percent of each panel and will be identified based on their background or experience in writing as well as with grade 4 students.	I assume this 30% target is based on previous SS sessions. I view this as being very high. For sure, this is a "public" activity, and for sure consistency with previous efforts is importnat. However, I just don't think that standards should be set by a group composed of 1/3 panelists who - no matter how bright, attentive, and knowledgeable about the construct of writing - have essentially no experience with 9 year olds and their writing.
11	Each panelist pool will include at least one teacher of English Language Learners and one teacher of students in Special Education programs;	This is the first recommendation that I am bothered by. Only 1 of 33 panelists will have EL experience, and 1 with SE?? THis does not strike me as advisable. I'd think on a panel of 33, you'd have 2 of each at a minimum. As you've outlined it, you'll likely have more (non-educator) parents or local school board members, or librarians than you will specialists in these key groups. I strongly urge you to reconsider this.
11	For the field trial, panelists will be drawn from a local pool of educators and non-educators who live within 30 miles of the field trial facility. This may limit Pearson's ability to obtain the broad representation intended on other panel characteristics, however, every effort will be made to ensure the representativeness and quality of the field trial panelists.	I don't see any reason (other than saving a very few dollars of travel cost) to restrict this so narrowly. Why not 50 or 60 miles - within an hour's drive??
14	Panelists nominated in each panelist group must meet the following qualifications:	minimal
15	We will also attempt to draw panels so that at least 20 percent of the panelists self-identify as a minority.	THis seems like a "throw in" here. I'd hope it would be a more-important consideration in your selection process
15	As noted above, in addition to covering the direct expenses for panelists (consistent with federal travel regulations), panelists for the field trial, pilot study, and operational ALS meeting will be given an honorarium of $500 each to cover incidental expenses during their stay at the panel meetings.	I'd hope that panelists don't really incur unreimbursed "incidental meeting-related expenses" Isn't the honorarium a payment, not a reimbursement???
16	Documents will include the following: Hotel information, including directions	INSERT: Travel and corresponding hotel information . . .

Page	Text/Reference	Comment
20	Based on findings of the studies conducted in 2011 for NAEP writing at grades 8 and 12 (Measured Progress, 2012a), the BoWs will be presented in descending order according to student performance based on the 2012 grade 4 NAEP pilot study.	I don't really object to arraying the work from high to low, but wonder what you mean here about "based on findings" from 2011. Did MP really find that high to low presentation had a material positive effect on the outcome? If so, I'm curious how they determined "better" - never mind, that's a topic for another day!
20	Other educators, composing about 15 percent of each panel, will include individuals such as higher education faculty teaching elementary education courses, librarians, and state and local English language arts curriculum directors.	why all the redundancy here??
21	The pre-meeting panelist briefing materials will include:	ditto - why necessary to repeat all of this??
21	The facilitator will lead the field trial meeting and guide panelist	Add an s to panelist
22	This question will include an open-ended component, asking panelists to explain their response—including the impact of the evidence presented on that response.	I'm not sure what you mean in this bullet, nor what you are hoping to gain from doing it. Maybe it's ok - even "good" - but I'm not sure that what a, say, curriculum director or a local school board member thinks about Grade 4 students' use of tablets is of any possible relevance here.
22	The facilitator will lead a short discussion of panelists' concerns if, after reviewing the relevant research, panelists continue to believe grade 4 students cannot use a tablet for writing.	Is this the point? YOu're trying to see is a panelist thinks that what you just proved could be done could be done? If they don't think it's possible, will you drop them or just check how their recommendations come out??
23	Facilitator providing an abbreviated orientation to and training on the grade 4 writing ALDs.	why "abbreviated"?? This is a crucial activity and I wouldn't be willing to assume that all the panelists internalized the essential ALD info from their at-home preparation.
24	Panelist selection of exemplars for at least one achievement level.	I assume you'll start here with Proficient, as it's the most important level??
30	As with the field trial, for all rounds of the ALS process, the BoWs will be presented in descending order according to student performance based on the 2017 grade 4 NAEP operational assessment.	I assume this is the procedure used for the earlier grades of Writing? Otherwise, it seems more logical to me that the BOWs would be sequenced from weak to strong as the typical SS procedure presents "items" from easy to hard. I bow to earlier procedures used

Page	Text/Reference	Comment
32	This training segment will include providing panelists with an understanding of the grade 4 NAEP writing ALDs, grounding them in what students should know and be able to do for the grade 4 NAEP writing assessment, and will conclude with a brief overview of the meeting schedule of activities.	I muse be misunderstanding this. You're going to overview the agenda after doing all of the above activities??
33	After submission of panelists' practice judgments, Pearson will review and analyze these judgments to provide panelists with appropriate feedback.	I'm confident that this process will be well thought-out. However, it seems advisable to address in more detail what you mean by this sentence - what will you DO, what will "appropriate feedback" entail, will you be attempting to shape panelist judgments or is this simply a "mechanical" process to ensure proper recording of judgments, etc. Anything that hints at altering the independence of panelist judgments should be explained in more detail than this single sentence.
34	Panelists will review each BoW and classify the set of responses as below Basic, Basic, Proficient, or Advanced.	I guess that "set" could mean "two," but this seems a bit overstated!
34	Panelists will review each BoW and classify the set of responses as below Basic, Basic, Proficient, or Advanced.	B (Capitalize)
34	In place of the pinpointing procedure, Round 3 will have panelists rate a new set of student responses, adding a replication component to the ALS process and providing further evidence for the evaluation of validity of ALS outcomes.	No problem with this "modification." (or the others undiscussed here) However, up to this point - page 34 of the document - I had been assuming you were proposing to use the "real" (not "modified") Body of Work methodology. I'd strongly urge you to add the modifier to earlier descriptions or somehow otherwise indicate to readers that you were making (what to me seems to be a fairly significant, though acceptable) modification to the generally understood process.
34	Logistic regression will be used to calculate cut scores based on panelists' ratings of each body of work into an achievement level (below Basic, Basic, Proficient, or Advanced).	why uncapitalized??
34	This curve then represents relationship among the EAP scores and panelist's ratings for all BoWs evaluated.	Add "the" (represents "the" relationship)

Page	Text/Reference	Comment
34	They will then discuss them as a table group in regard to the skills required in the ALDs and represented in the BoWs..	Remove extra period
34	The facilitator will then lead a whole group	whole-group
35	The BoWs will be ordered from the highest to lowest performance, based on EAP score estimates.	Need to state here how the 50 BOWs will be chosen to range across the 11-point score scale. That is, will you choose 4-5 BOWs per score point, underweight the extremes, overweight the middles, overweight anticipated cutscores, . . . This is CRITICAL
36	Had a reversal (for example, lower scoring bodies of work were classified into higher achievement levels than higher scoring bodies of work)	I had assumed from earlier discussion that the panelists know that they were viewing the BOWs in score-point order. THus, isn't it pretty unlikely that they'd choose to place a higher-scoring product into a lower achievement level??
36	These data will show the percentage of students at or above each median panel cut score, using the results of the 2017 grade 4 NAEP writing assessment.	Will you again show the higher-grade impact data??
37	Each panelist will then use the feedback from Round 2, as well as their discussions with other panelists, to provide judgments for a new, but comparable, set of BoWs.	I'm not sure I see how "feedback" from an unrelated set of papers is supposed to be helpful here.
45	Document such as the use of a computer based assessment of writing for fourth grade students and the timing and content of the advanced materials.	advance
45	Cizek, G. J., Bunch, B. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. [An NCME instructional module]. *Educational Measurement: Issues and Practice, 23*(4), 31–50.	change first “B.” to “M.”

A White Paper on Uses and Misuses of NAEP Data

On the recommendation of the Governing Board, NCES requested that the NAEP Validity Studies Panel (NVS) develop a white paper that is intended to inspire a short and modest statement of intended uses and interpretations of NAEP scores. This request was based on the observation that the first standard from the AERA/APA/NCME standards is, "…the test developer should set forth clearly how test scores are intended to be interpreted and consequently used." The paper will discuss a variety of intended interpretations and uses that are explicit and implicit across NAEP legislation and products. It will also discuss interpretations and uses that may not be intended but are nonetheless widespread.

Finally, the paper will review misuses of NAEP data including, but not limited to, comparisons among subscales and subjects, faulty (causal) inferences, and so on. It will conclude with a discussion of the kind of evidentiary bases needed for valid uses and interpretations of NAEP data.

In this COSDAM session, Dr. Ina Mullis of the NAEP Validity Studies Panel will present plans for the white paper.

Follow Up to Evaluation of NAEP Achievement Levels

On November 17, 2016, the National Academies of Sciences, Engineering, and Medicine released the final report of their evaluation, *Evaluation of the achievement levels for mathematics and reading on the National Assessment of Educational Progress*. A free PDF of the full report can be downloaded at: https://www.nap.edu/catalog/23409/evaluation-of-the-achievement-levels-for-mathematics-and-reading-on-the-national-assessment-of-educational-progress. The Governing Board received a briefing from staff at the National Academies of Sciences, Engineering, and Medicine and members of the interdisciplinary review committee during the most recent quarterly Board meeting on November 19, 2016.

As stated in the NAEP legislation, the Commissioner of the National Center for Education Statistics (NCES) is to use the findings from the evaluation to decide whether the achievement levels should continue to be used on a "trial basis" or whether that designation can be removed. In addition, the final report included conclusions and recommendations that have implications for future Governing Board achievement levels-setting activities. Public Law 107-279 specifies that the Governing Board must prepare a formal response to the evaluation:

> *Not later than 90 days after an evaluation of the student achievement levels under section 303(e), the Assessment Board shall make a report to the Secretary, the Committee on Education and the Workforce of the House of Representatives, and the Committee on Health, Education, Labor, and Pensions of the Senate describing the steps the Assessment Board is taking to respond to each of the recommendations contained in such evaluation.*

Due to the timing of the evaluation report release, the 90 day window concluded prior to the March 2017 Governing Board meeting. Therefore, on November 19, 2016, the Board granted a joint delegation of authority to COSDAM and the Executive Committee for formal approval of the report to the Secretary, the Committee on Education and the Workforce of the House of Representatives, and the Committee on Health, Education, Labor, and Pensions of the Senate describing the steps the Governing Board is taking to respond to each of the recommendations contained in the evaluation.

COSDAM met via teleconference on December 9, 2016 to discuss an initial draft response to the evaluation. On December 19, 2016, the Executive Committee and COSDAM met to discuss and take action on a revised response. The final response was approved by a vote of 9-0 with one abstention. The response was sent to Secretary John King, the Committee on Education and the Workforce of the House of Representatives, and the Committee on Health, Education, Labor, and Pensions of the Senate on December 20, 2016.

The Governing Board response refers to several new activities to be undertaken, in addition to plans to update the current policy on Developing Student Performance Levels for NAEP (attached). Much of the work aligns with the Strategic Vision and can be performed in

collaboration with NCES. Ongoing discussions with COSDAM and the full Board will take place over the next several quarterly meetings to plan and implement the recommendations from the evaluation.

During the March 2, 2017 meeting, COSDAM members will discuss the Governing Board response, including next steps and priorities related to new procurements and updates to the current Governing Board policy on achievement levels setting.

National Assessment Governing Board's Response to the National Academies of Sciences, Engineering, and Medicine 2016 Evaluation of NAEP Achievement Levels

Legislative Authority

Pursuant to the National Assessment of Educational Progress (NAEP) legislation (Public Law 107-279), the National Assessment Governing Board (hereafter the Governing Board) is pleased to have this opportunity to apprise the Secretary of Education and the Congress of the Governing Board response to the recommendations of the National Academies of Sciences, Engineering, and Medicine evaluation of the NAEP achievement levels for mathematics and reading (Edley & Koenig, 2016).

The cited legislation charges the Governing Board with the authority and responsibility to "develop appropriate student achievement levels for each grade or age in each subject area to be tested." The legislation also states that "such levels shall be determined by... a national consensus approach; used on a trial basis until the Commissioner for Education Statistics determines, as a result of an evaluation under subsection (f), that such levels are reasonable, valid, and informative to the public; ... [and] shall be updated as appropriate by the National Assessment Governing Board in consultation with the Commissioner for Education Statistics" (Public Law 107-279).

Background

NAEP is the largest nationally representative and continuing assessment of what our nation's elementary and secondary students know and can do. Since 1969, NAEP has been the country's foremost resource for measuring student progress and identifying differences in student achievement across student subgroups. In a time of changing state standards and assessments, NAEP serves as a trusted resource for parents, teachers, principals, policymakers, and researchers to compare student achievement across states and select large urban districts. NAEP results allow the nation to understand where more work must be done to improve learning among all students.

For 25 years, the NAEP achievement levels (*Basic*, *Proficient*, and *Advanced*) have been a signature feature of NAEP results. While scale scores provide information about student achievement over time and across student groups, achievement levels reflect the extent to which student performance is "good enough," in each subject and grade, relative to aspirational goals. Since the Governing Board began setting standards in the early 1990s, achievement levels have become a standard part of score reporting for many other assessment programs in the US and abroad.

Governing Board Response

Overview

The Governing Board appreciates the thorough, deliberative process undertaken over the past two years by the National Academies of Science, Engineering, and Medicine and the expert members of the Committee on the Evaluation of NAEP Achievement Levels for Mathematics and Reading. The Governing Board is pleased that the report concludes that the achievement levels are a meaningful and important part of NAEP reporting. The report states that, "during their 24 years [the achievement levels] have acquired meaning for NAEP's various audiences and stakeholders; they serve as stable benchmarks for monitoring achievement trends, and they are widely used to inform public discourse and policy decisions. Users regard them as a regular, permanent feature of the NAEP reports" (Edley & Koenig, 2016; page Sum-8). The Governing Board has reviewed the seven recommendations presented in the report and finds them reasonable and thoughtful. The report will inform the Board's future efforts to set achievement levels and communicate the meaning of NAEP *Basic*, *Proficient*, and *Advanced*. The recommendations intersect with two Governing Board documents, the Strategic Vision and the achievement levels policy, described here.

On November 18, 2016, the Governing Board adopted a Strategic Vision (https://www.nagb.org/content/nagb/assets/documents/newsroom/press-releases/2016/nagb-strategic-vision.pdf) to guide the work of the Board through 2020, with an emphasis on innovating to enhance NAEP's form and content and expanding NAEP's dissemination and use. The Strategic Vision answers the question, "How can NAEP provide information about how our students are doing in the most innovative, informative, and impactful ways?" The Governing Board is pleased that several of the report recommendations are consistent with the Board's own vision. The Governing Board is committed to measuring the progress of our nation's students toward their acquisition of academic knowledge, skills, and abilities relevant to this contemporary era.

The Governing Board's approach to setting achievement levels is articulated in a policy statement, "Developing Student Performance Levels for the National Assessment of Educational Progress" (https://www.nagb.org/content/nagb/assets/documents/policies/developing-student-performance.pdf). The policy was first adopted in 1990 and was subsequently revised in 1995, with minor wording changes made in 2007. The report motivates the revision of this policy, to add clarity and intentionality to the setting and communication of NAEP achievement levels.

The seven recommendations and the Governing Board response comprise a significant research and outreach trajectory that the Governing Board can pursue over several years in conjunction

with key partners. The Governing Board will implement these responses within resource constraints and in conjunction with the priorities of the Strategic Vision.

Evaluating the Alignment of NAEP Achievement Level Descriptors

Recommendation #1: Alignment among the frameworks, the item pools, the achievement-level descriptors, and the cut scores is fundamental to the validity of inferences about student achievement. In 2009, alignment was evaluated for all grades in reading and for grade 12 in mathematics, and changes were made to the achievement-level descriptors, as needed. Similar research is needed to evaluate alignment for the grade 4 and grade 8 mathematics assessments and to revise them as needed to ensure that they represent the knowledge and skills of students at each achievement level. Moreover, additional work to verify alignment for grade 4 reading and grade 12 mathematics is needed.

The report's primary recommendation is to evaluate the alignment, and revise if needed, the achievement level descriptors for NAEP mathematics and reading assessments in grades 4, 8, and 12. The Governing Board intends to issue a procurement for conducting studies to achieve this goal. The Governing Board has periodically conducted studies to evaluate whether the achievement level descriptors in a given subject should be revised, based on their alignment with the NAEP framework, item pool, and cut scores. The Governing Board agrees that this is a good time to ensure that current NAEP mathematics and reading achievement level descriptors align with the knowledge and skills of students in each achievement level category. In conjunction with the response to Recommendation #3, the updated Board policy on NAEP achievement levels will address the larger issue of specifying a process and timeline for conducting regular recurring reviews of the achievement level descriptions in all subjects and grades.

The Governing Board agrees strongly with the recommendation that, while evaluating alignment of achievement level descriptors is timely, it is not necessary to consider changing the cut scores or beginning a new trend line at this time. The NAEP assessments are transitioning from paper-based to digital assessments in 2017, and current efforts are focused on ensuring comparability between 2015 and 2017 scores. The Governing Board articulated this in the 2015 Resolution on Maintaining NAEP Trends with the Transition to Digital-Based Assessments (https://www.nagb.org/content/nagb/assets/documents/policies/resolution-on-trend-and-dba.pdf).

Recommendation #2: Once satisfactory alignment among the frameworks, the item pools, the achievement-level descriptors, and the cut scores in NAEP mathematics and reading has been demonstrated, their designation as trial should be discontinued. This work should be completed and the results evaluated as stipulated by law: (20 U.S. Code 9622: National Assessment of Educational Progress: https://www.law.cornell.edu/uscode/text/20/9622 [September 2016]).

Ultimately, the Commissioner of Education Statistics is responsible for determining whether the "trial" designation is removed. The Governing Board is committed to providing the Commissioner with the information needed to make this determination in an expedient manner.

Regular Recurring Reviews of the Achievement Level Descriptors

Recommendation #3: To maintain the validity and usefulness of achievement levels, there should be regular recurring reviews of the achievement-level descriptors, with updates as needed, to ensure they reflect both the frameworks and the incorporation of those frameworks in NAEP assessments.

The Board's current policy on NAEP achievement levels contains several principles and guidelines for *setting* achievement levels but does not address issues related to the continued use or reporting of achievement levels many years after they were established. The revised policy will seek to address this gap by including a statement of periodicity for conducting regular recurring reviews of the achievement level descriptors, with updates as needed, as called for in this recommendation. The Governing Board agrees that it is important to articulate a process and timeline for conducting regular reviews of the achievement level descriptors rather than performing such reviews on an ad hoc basis.

Relationships Between NAEP Achievement Levels and External Measures

Recommendation #4: Research is needed on the relationships between the NAEP achievement levels and concurrent or future performance on measures external to NAEP. Like the research that led to setting scale scores that represent academic preparedness for college, new research should focus on other measures of future performance, such as being on track for a college-ready high school diploma for 8th-grade students and readiness for middle school for 4th-grade students.

In addition to the extensive work that the Governing Board has conducted at grade 12 to relate NAEP mathematics and reading results to academic preparedness for college, the Governing Board has begun research at grade 8 with statistical linking studies of NAEP mathematics and reading and the ACT Explore assessments in those subjects. This work was published while the evaluation was in process and was not included in the Committee's deliberations. Additional studies in NAEP mathematics and reading at grades 4 and 8 are beginning under contract to the National Center for Education Statistics (NCES). The Governing Board's Strategic Vision includes an explicit goal to increase opportunities for connecting NAEP to other national and international assessments and data. Just as the Board's previous research related grade 12 NAEP results in mathematics and reading to students' academic preparedness for college, the Governing Board anticipates that additional linkages with external measures will help connect the NAEP achievement levels and scale scores to other meaningful real-world indicators of current and future performance.

Interpretations and Uses of NAEP Achievement Levels

Recommendation #5: Research is needed to articulate the intended interpretations and uses of the achievement levels and collect validity evidence to support these interpretations and uses. In addition, research to identify the actual interpretations and uses commonly made by NAEP's various audiences and evaluate the validity of each of them. This information should be communicated to users with clear guidance on substantiated and unsubstantiated interpretations.

The Governing Board's Strategic Vision emphasizes improving the use and dissemination of NAEP results, and the Board's work in this area will include achievement levels. The Governing Board recognizes that clarity and meaning of NAEP achievement levels (and scale scores) are of utmost importance. The Governing Board will issue a procurement to conduct research to better understand how various audiences have used and interpreted NAEP results (including achievement levels). The Governing Board will work collaboratively with NCES to provide further guidance and outreach about appropriate and inappropriate uses of NAEP achievement levels.

Guidance for Inferences Made with Achievement Levels versus Scale Scores

Recommendation #6: Guidance is needed to help users determine inferences that are best made with achievement levels and those best made with scale score statistics. Such guidance should be incorporated in every report that includes achievement levels.

The Governing Board understands that improper uses of achievement level statistics are widespread in the public domain and extend far beyond the use of NAEP data. Reports by the Governing Board and NCES have modeled appropriate use of NAEP data and will continue to do so. This recommendation is also consistent with the goal of the Strategic Vision to improve the dissemination and use of NAEP results. The Governing Board will continue to work with NCES and follow current research to provide guidance about inferences that are best made with achievement levels and those best made with scale score statistics.

Regular Cycle for Considering Desirability of Conducting a New Standard Setting

Recommendation #7: NAEP should implement a regular cycle for considering the desirability of conducting a new standard setting. Factors to consider include, but are not limited to: substantive changes in the constructs, item types, or frameworks; innovations in the modality for administering assessments; advances in standard setting methodologies; and changes in the policy environment for using NAEP results. These factors should be weighed against the downsides of interrupting the trend data and information.

When the Board's achievement levels policy was first created and revised in the 1990s, the Board was setting standards in each subject and grade for the first time and had not yet considered the need or timeline for re-setting standards. To address this recommendation, the Governing Board will update the policy to be more explicit about conditions that require a new standard setting.

Board's Commitment

The Governing Board remains committed to its congressional mandate to set "appropriate student achievement levels" for the National Assessment of Educational Progress. The Board appreciates the report's affirmation that NAEP achievement levels have been set thoughtfully and carefully, consistent with professional guidelines for standard setting, and based on extensive technical advice from respected psychometricians and measurement specialists. The Board also takes seriously the charge to develop the current achievement levels through a national consensus approach, involving large numbers of knowledgeable teachers, curriculum specialists, business leaders, and members of the general public throughout the process. This is only fitting given the Governing Board's own congressionally mandated membership that explicitly includes representatives from these stakeholder groups.

The Governing Board remains committed to improving the process of setting and communicating achievement levels. The Governing Board is grateful for the report recommendations that will advance these aims.

Reference

Edley, C. & Koenig, J. A. (Ed.). (2016). *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress.* Washington, DC: The National Academies Press.

Adopted: March 4, 1995

National Assessment Governing Board

Developing Student Performance Levels for the National Assessment of Educational Progress

Policy Statement

Foreword
A policy on setting achievement levels on the National Assessment of Educational Progress (NAEP) was first adopted in 1990 and amended several times thereafter. The present policy, adopted in 1995, contained introductory and explanatory text, principles, and guidelines. Since 1995, there have been several changes to the NAEP authorizing legislation (currently, the NAEP Authorization Act: P.L. 110-279). In addition, related legislation has been enacted, including the No Child Left Act of 2001. Consequently, introductory and other explanatory text in the original version of this policy, no longer germane, has been deleted or revised to conform to current legislation. The Principles and Guidelines remain in their original form except for Principle 4, from which the reference to the now decommissioned Advisory Council on Education Statistics has been deleted. (Foreword added August 2007.)

Principles for Setting Achievement Levels

Principle 1

The level setting process shall produce for each content area, three threshold points at each grade level assessed, demarcating entry into three categories: *Basic, Proficient, and Advanced.*

Proficient. *This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.*

Basic.	*This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.*
Advanced.	*This level signifies superior performance beyond proficient.*

Principle 2

Developing achievement levels shall be a widely inclusive activity of the Board, utilizing a national consensus approach, and providing for the active participation of teachers, other educators (including curriculum specialists and school administrators at the local and state levels), and non-educators including parents, members of the general public, and specialists in the particular content area.

The development of achievement levels shall be conducted in two phases. In phase 1, the assessment framework development process shall yield preliminary descriptions of the achievement levels *(Basic, Proficient, and Advanced),* which shall subsequently be used in phase 2 to develop the numerical standards (cut scores) and to identify appropriate examples of assessment exercises that typify performance at each level. The levels will be updated as appropriate, typically when the assessment frameworks are updated.

Principle 3

The Governing Board shall incorporate the student performance levels into all significant elements of NAEP, including the subject area framework development process, exercise development and selection, and the methodology of the assessment. The achievement levels shall be used to report the results of the NAEP assessments so long as such levels are reasonable, valid and informative to the public.

Principle 4

In carrying out its statutory mandate, the Governing Board will *exercise its policy judgment in setting the levels.* The Board shall continually seek better means of setting achievement levels. In so doing, the Board may seek technical advice as appropriate from a variety of sources, including external evaluations provided by the Secretary, the Commissioner, and other experts. Proposed achievement levels shall be reviewed by a broad constituency, including consumers of NAEP data, such as policymakers, professional groups, the states and territories. In carrying out its responsibilities, the Board will ordinarily engage the services of a contractor who will prepare recommendations for the Board's consideration on the levels, the descriptions, and the exemplar exercises.

Guidelines for Setting Achievement Levels

Each guideline presented below is accompanied by a rationale and a summary of the implementation practices and procedures to be followed in carrying out the principle. It should be understood that the full implementation of this policy will require the

contractor, through Governing Board staff, to provide assurances to the Board that all aspects of the practices and procedures for which they are responsible have been completed successfully. These assurances will be in writing, and may require supporting documentation prepared by the contractor and/or Governing Board staff.

Summary of Guidelines

Guideline 1

The level setting process shall produce for each content area, three threshold points at each grade level assessed, demarcating entry into three categories: *Basic, Proficient, and Advanced.*

Guideline 2

The level setting process shall be a widely inclusive activity of the Board, carried out by a broadly representative body of teachers, other educators (including curriculum specialists and local and state administrators), and non-educators including parents, concerned members of the general public, and specialists in the particular content area; this process and resulting products shall be reviewed by a broad constituency.

Guideline 3

The level-setting process shall result in achievement level cut scores for each grade and level, expanded descriptions of the content expected at each level based on the preliminary descriptions provided through the national consensus process, and exemplar exercises that are representative of the performance of examinees at each of the levels and of the cognitive expectations for each level described.

Guideline 4

In carrying out its statutory mandate, the Board will *exercise its policy judgment in setting the levels.* However, in so doing, they will seek technical advice from a variety of sources, but especially from the contractor who will prepare the recommendations on the levels, the descriptions, and the exemplar exercises, as well as from consumers of NAEP data, including policymakers, professional groups, the states, and territories.

Guideline 5

The achievement levels shall be the initial and primary means of reporting the results of the National Assessment of Educational Progress at both the national and state levels.

Guideline 6

The level-setting process shall be managed in a technically sound, efficient, cost-effective manner, and shall be completed in a timely fashion.

Guideline 1

The level setting process shall produce for each content area, three threshold points at each grade level assessed, demarcating entry into three categories: *Basic, Proficient, and Advanced.*

Rationale

The Board is committed to describing the full range of performance on the NAEP scale, for students whose performance is in the mid-range, as well as for those whose performance is below and above the middle. It is highly desirable to endorse realistic expectations for all students to achieve no matter what their present performance might be. Three benchmarks on the NAEP scale suggest realistic expectations for students in all regions of the performance distribution. Likewise, the Board is committed to preserving trend results in NAEP. Three achievement levels accommodate growth (and possible declines) in all ranges of the performance distribution.

Practices and Procedures

Policy Definitions

The following policy definitions will be applied to all grades, 4, 8, and 12, and all content areas in which the levels are set. It is the Board's view that the level of performance referred to in the policy definitions is what students *should be able to know and do,* and not simply the current academic achievement of students or that which today's U.S. schools expect.

Proficient.	*This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.*
Basic.	*This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.*
Advanced.	*This level signifies superior performance beyond proficient.*

From Policy Definitions to Content Descriptions

In the course of applying the policy definitions to the level-setting process, it will be necessary to articulate them in terms of the specific content and sequence (now called descriptions) appropriate for the grades in which the levels are being set. This will be completed on a preliminary basis through the process which develops the assessment

frameworks. These preliminary descriptions will be used to initially guide the work of deriving the advice that will assist the Board in setting the levels. Throughout the process of obtaining such advice, however, these descriptions may be refined, expanded, and edited to more clearly reflect the specific advice on the levels.

Training of Judges

In training the judges for the level-setting activity, it is necessary that all arrive at a common conceptualization of *Basic, Proficient, and Advanced* based on the policy definitions of the Board. Such conceptualizations must be within the scope of the assessment framework under consideration and capable of being applied at the individual item level (Reid, 1991.)

Judges must also be trained in the specific model that will be used to generate the rating data. At the very least, they need to understand the purposes for setting the levels, the significance of such an activity, the NAEP assessment framework for the subject area under discussion, elements that make particular exercises more or less difficult, and the rating task itself.

Judges shall be trained by individuals who are both knowledgeable in the subject matter area and are experienced, capable trainers in a large-group setting. Presentations shall be prepared, rehearsed, and piloted before implementation.

Judges shall be provided comprehensive, user-friendly training materials, adequate time to complete the task, and the appropriate atmosphere in which to work, one that is quiet, pleasant, and conducive to reaching the goals of the level-setting activity. It is also required that judges take the assessment under the same NAEP-like conditions as students, that is, using the NAEP student booklets, having all manipulatives and ancillary materials, and timed.

Guideline 2

The level setting process shall be a widely *inclusive* activity of the Board, carried out by a broadly representative body of teachers, other educators (including curriculum specialists and local and state administrators), and non-educators including parents, concerned members of the general public, employers, scholars, and specialists in the particular content area. This process and resulting products shall be reviewed by a broad constituency.

Rationale

The spirit of the legislative mandate of the Board is one of moving toward a national consensus on policy issues affecting NAEP. The Board has historically involved broad audiences in its deliberations. The achievement levels are no different. Further, the Board views the level-setting activity as an extension of the widely inclusive effort to derive the assessment frameworks and scope and sequence of each assessment. Finally, the magnitude of the decisions regarding *what students should know and be able to do is*

simply too important a decision to seek involvement from professionals alone; it must have the benefit of the collective wisdom of a broadly representative body, educators and non-educators alike.

Practices and Procedures

Sample of Judges

The panel of judges will be composed of both educators and non-educators. About two-thirds of the panel will represent teachers and other educators; one-third will represent the public, non-educator sector, for example, scholars, employers, parents, and professionals in occupations related to the content area. They will be drawn from a national sampling frame and will be broadly representative of various geographic regions (Northeast, Southeast, Central, West, and the territories) types of communities (urban, suburban, rural), ethnicities, and genders.

Individual panel members shall have expertise in the specific content area in which the levels are being developed, expertise in the education of students at the grades under consideration, and a general knowledge of assessment, curriculum, and student performance. The composition of the panels should be such that they meet the requirements of the *Standards (1985).*

The size of the panels should be responsive to what the research demonstrates regarding numbers of judges involved (see Jaeger, 1991). While it may not be practical or beyond the resources available, every effort should be made to empanel a sufficient number of judges to reduce the standard error of the cut score. While there is no absolute criterion on the magnitude of the standard error of the cut score, a useful rule of thumb is that it should not exceed the *combined* error associated with the standard error of measurement on the assessment and the error due to sampling from the population of examinees.

Review Procedures

Throughout the process and particularly at critical junctures, groups that have a legitimate interest in the process will be involved. During the planning process interested groups and individuals will be encouraged to participate and share their experiences in the area of setting standards. These groups might include professional societies, *ad hoc* advisory groups, standing advisory committees to the Governing Board or its contractor(s) and NCES and its contractor(s) and grantees. Documents (such as the Design Document and Interim Reports) will be disseminated in sufficient time to allow for a thoughtful response from those who wish to provide one.

Proposed levels will be widely distributed to major professional organizations, state and local assessment and curriculum personnel, business leaders, government officials, the Planning and Steering Committees of the framework development process, the Exercise Development panels, and other groups who may request them.

When it is deemed useful by the Board, public hearings and forums will be conducted in Washington, D.C. and other parts of the country to encourage review and input on a broad regional and geographic basis.

Guideline 3

The resulting products of the level-setting process shall be (1) achievement level scores marking the threshold score for each grade and level, (2) expanded descriptions of the content expected at each level based on the preliminary descriptions provided through the national consensus process, and (3) exemplar exercises that are representative of the performance of examinees at each of the levels and of the cognitive expectations for each level described. These three products form the basis for reporting the results of all future NAEP assessments.

Rationale

The NAEP scale, while useful for aggregating large amounts of information about student performance in a single number, requires contextual information about the specific content and the sequencing of that content across particular grades, in order to be truly beneficial to users of NAEP data. In order to make the NAEP data more useful, descriptions of each level which articulate content expectations and exemplar exercises taken from the public release pool of the most current NAEP assessment must accompany the benchmarks or cut scores for each level. The descriptions and exemplars are intended to be illustrative of the kind of content that is represented in the levels, as well as an aid in the interpretation of the NAEP data.

Practices and Procedures

Methodology

The methodology to be used in generating the levels will depend upon the specific assessment formats for the content area in which the levels are being set. Historically, in the case of multiple choice exercises and short constructed response formats, a modified Angoff (1971) procedure has been employed. In the case of extended constructed response formats, a paper-selection procedure has been employed. Neither of these is without its disadvantages. As the assessment formats of future assessments become more complex and employ more performance-type exercises, it is quite likely that alternate procedures will be needed. The Board will decide these on a case-by-case basis, looking for advice from those who have had experience in dealing with these alternative assessment formats. In any case, the design for carrying out the process must be carefully crafted, must be appropriate to the content area and philosophy of the assessment framework, and must have a solid research base.

The procedures will generally be piloted prior to full implementation. The purpose of the pilot would be to test out the materials used with the judges, the training procedures, the feedback information given to the judges during the process, and the

software used to complete the initial analyses. Procedures would be revised based on the pilot experience and evaluation evidence.

Whatever methodology is used, all aspects of the procedures will be documented for the purposes of providing evidence of procedural validity for the levels being recommended. This evidence will be made available to the Board at the time of deliberations about the levels being set.

Quality Control Procedures

While there are numerous points in a complex process for mistakes to occur, there are at least three important junctures where quality control measures need to be in place. First, is the point of data entry. Ideally, judges' ratings should be scanned to reduce manual errors of entry. However, if the ratings are entered manually, then they shall be entered and 100% verified using a double-entry, cross-checking procedure. Second, software programs designed to complete initial analyses on the rating data must be run with simulated data to de-bug, and provide assurances of quality control. The programs should detect logical errors and other kinds of problems that could result in incorrect results being generated. Finally, the production of cut scores on the NAEP scale is the final responsibility of the NAEP operations contractor. Only final cut scores, mapped onto the properly weighted and equated scale, received in writing from the operations contractor, will be officially communicated to the Board, or others who have a legitimate need to know. *Once the accuracy of the data has been ensured by the level-setting and operations contractors, the Board shall make a policy determination and set the final achievement levels, informed by the technical process of the level-setting activity.*

Descriptions of the Levels

The preliminary descriptions developed through the framework development process will be the starting point for developing recommendations for the levels under consideration. The preliminary descriptions are *working descriptions* for the panels while doing the ratings. These may be expanded and revised accordingly as these panels conduct the ratings, examine empirical performance data, and work to develop their final recommendations on the levels. The recommended descriptions will be articulated in terms of what students *should know and should be able to do.* They shall be coherent within grade, and consistent across grades, and will reference performance within the three regions created by the cut scores. No descriptions will be done for content below the *Basic* level.

Exemplar Exercises

The exemplars chosen from the released pool of exercises for the current NAEP assessment will reflect as much as possible performance both in the *Basic, Proficient, and Advanced* regions of the scale, as we1l as at the threshold scores. Exemplars will be selected to meet the rp = *.50* criterion, and will demonstrate the range of performance possible within the regions. They will likewise reflect the content found in the final descriptions and the range of item formats on the assessment. Evidence will be provided for the degree of congruence between the content of the exemplars and that of the descriptions. There will be at least three exemplars per level per grade identified.

Guideline 4

In carrying out its statutory mandate, the Board will *exercise its policy judgment in setting the levels.* However, in so doing, they will seek technical advice from a variety of sources, but especially from the contractor, who will prepare the recommendations on the levels, the descriptions, and the exemplar exercises, as well as from consumers of NAEP data, including policymakers, professional groups, the states and territories.

Rationale

Setting achievement levels is both an *art* and a *science.* As an *art,* it requires judgment. It is the Board's best policy judgment what the levels should be. However, as a *science,* it requires solid technical advice based on a sound technical process. The Board is committed to seeking such technical advice from a variety of sources.

Practices and Procedures

Technical Advice throughout the Process

The Board seeks to involve persons who have had experience in standard-setting at the state level, and from those who are users of the NAEP results. Regular presentations will be given to standing committees who advise on NAEP matters such as the Education and Information Advisory Committee (EIAC) of the CCSSO, and the NAEP NETWORK. Their counsel will be sought on matters of substance as the work of the Board progresses. The EIAC and other similar constituencies may also be invited to send a representative to all standing technical advisory committees of the Board's contractor(s) which deal with the level-setting process.

The Board will also seek advice from the technical community throughout the level-setting process. Efforts will be made to ensure that presentations are made regularly to such groups as the American Educational Research Association (AERA), the National Council for Measurement in Education (NCME), and the professional groups in the content areas such as the International Reading Association (IRA), the National Science Teachers Association (NSTA), and other similar organizations. The Board will seek to engage technical groups available to them, including the Technical Review Panel, the National Academy of Education, their own contractor(s), and NCES and its contractor(s), in constructive research studies focused on providing information on the technical aspects of NAEP related to level-setting (e.g., scaling, weighting, mapping ratings to the scale, etc.)

Validity and Reliability Evidence

The Board will examine and consider all evidence of reliability and validity available. These data would include, but need not be limited to, procedural evidence such as the selection and training of judges and the materials and methods used in the process, reliability evidence such as intra-judge and inter-judge consistency data, and finally, internal and external validity data. Such data will help to inform *the Board's policy decision as they set the levels.*

Procedural evidence, while informative, is not necessarily sufficient evidence for demonstrating the validity of the levels. Therefore, the conduct of the achievement level-setting process shall be implemented so that a series of both internal and external validation studies shall be conducted simultaneously. To the extent possible, in order to realize maximum efficiencies in the use of resources, validation studies shall be included in the design of the level-setting data collection activities. Such studies may include, but shall not be limited to, convergent and divergent validation efforts, for example, conducting alternate standard-setting methods or conducting cross-validation level-setting activities, as well as exploring alternate methods for refining and expanding the preliminary achievement levels definitions, and empirically examining various technical decision rules used throughout the process.

As part of the validation task, additional evidence as to the suitability and appropriateness of identifying the subject area content of the recommended achievement levels ranges and cut-scores will be gathered. This evidence may include, but need not be limited to, data resulting from behaviorally anchoring the ranges and/or cut-scores, or data resulting from some other alternative procedures that employ a more global approach other than the item content of the particular assessment. The results of these studies will provide a clear indication of what students know and can do at the levels.

The results from these validation efforts shall be made available to the Board in a timely manner so that the Board has access to as much validation data as possible as it considers the recommendations regarding the final levels. Kane (1993) suggests that an "interpretive argument would specify the network of inferences leading from the score to the conclusions drawn about examinees and the decisions made about examinees, as well as the assumptions that support these inferences." An interpretative argument which articulates the rationale for interpreting the levels shall accompany the presentation of proposed levels to the Board.

Again, to maximize the efficient use of resources and to minimize duplication of effort, it is highly desirable for contractors to coordinate the design of such studies with other agencies responsible for evaluating the level-setting activities.

Guideline 5

The achievement levels shall be the initial and primary means of reporting the results of the National Assessment of Educational Progress at both the national and state levels.

Rationale

In an effort to improve the form and use of NAEP the Board seeks to make the results of NAEP more accessible and understandable to the general public and to policy makers. The Board also supports the movement from norms-based assessments to standards-based assessments. Reporting the results of NAEP using the achievement levels accomplishes these ends to a greater degree than heretofore possible.

Practices and Procedures

Reporting What Students Know and Can Do

The purpose of most NAEP reports, but particularly those published under the auspices of the National Center for Education Statistics, is to report to the American public and others on the performance of students—that is, to report on *what students know and can do.* The purpose of the achievement levels is to identify for the American public what students *should know and should be able to do,* and to report the actual performance of students in relation to the achievement levels. Therefore, NAEP reports incorporate elements of both of these aspects of performance.

Clarity of interpretation of the NAEP data can be achieved by ensuring that the descriptions of performance for the levels and the exemplar exercises reflect what the empirical data show for a given assessment. This may be achieved by the modified procedures of *scale anchoring* [1] or by new procedures developed specifically for the purposes of providing elements of the content of the frameworks in the reporting mechanisms.

Reporting Student Performance

In describing student performance using the levels, terms such as *students performing at the Basic level* or *students performing at the Proficient level* are preferred over *Basic students* or *Proficient students.* The former implies that students have mastery of particular content represented by the levels, while the latter implies an inherent characteristic of individual students.

In reporting the results of NAEP, the application of the levels of *Basic, Proficient, and Advanced* applies to the three regions of the NAEP scale generated when the appropriate cut scores are mapped to the scale. However, three cut scores yield, in fact, four regions. The region referenced by content which falls below the *Basic* cut score will be identified by descriptors that are not value-laden.

Interpreting Student Performance

When interpreting student performance using the levels, one must diligently avoid over interpretations. For example, each of the NAEP subject areas are scaled independently of each other, even though each scale uses the same metric, i.e., scores ranging from 0 to 500. Because the metrics are identical, it does not follow that comparisons can be made across subjects. For example, a *Proficient* cut score of *235* in reading should not be interpreted to have the same meaning as a *Proficient* cut score of *235* in U.S. history. Neither should unwarranted comparisons be made in the same subject area from one assessment year to the next, unless the data for the two years have been equated and we have reason to believe that the scale itself has not changed from time 1 to time 2.

Guideline 6

The level-setting process shall be managed in a technically sound, efficient, cost-effective manner, and shall be completed in a timely fashion.

Rationale

Since a contractor(s) is conducting technical advisory and assistance work for the Board, it is critical that such work be performed to meet high quality standards, including efficiency, cost-effectiveness, timeliness, and adherence to sound measurement practices. *However, in the final analysis, it is the Governing Board that makes the policy decision regarding the levels, not the contractor.*

Practices and Procedures

The contractor(s) shall prepare a fully detailed Planning Document at the onset of the level-setting work. This document will guide the progress of the work, serve as a monitor, and be the basis for staff and Board supervision. The Planning Document will outline milestone events in the process, provide a chronology of tasks and subtasks, as well as a monthly chronology of all activities across all tasks, and detail all draft and final documents that will be produced, the audience for such reports, and the number of copies to be provided by the contractor.

Procedures adopted by a contractor(s) to carry out the level-setting process must encourage and support national involvement by the relevant and required publics. Such meetings will also be conducted in a physical environment which is conducive to work and planning. To the extent possible, current technology shall be used in all areas of the level-setting process to increase efficiency and to reduce error.

The contractor(s) shall work closely and in a professional manner with the NAEP operations contractor in striving to fulfill the requirements of the level-setting process by (1) making all requests for information and data in a timely manner, (2) providing all requested information and data in a timely manner, (3) adhering to all predetermined deadlines so as not to impede the work of the operations contractor, and (4) advising the operations contractor of all unusual findings in the data so that a concerted effort can be mounted to resolve the problem or issue at hand.

The contractor(s) shall develop the initial level-setting design adhering to sound measurement principles and ensure that the various components of the design (e.g., selection of judges) are congruent with current standard-setting research. In the implementation of such designs, they shall employ state-of-the-art training strategies and measurement practices.

The contractor(s) shall produce documents in a timely manner and make oral presentations upon request. Presentations may include, but need not be limited to, the Board's quarterly meetings, relevant Board committees, and professional and lay groups.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *(1985). Standards for Educational and Psychological Testing.* Washington, DC: APA.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement (2nd ed., pp. 508-600).* Washington, DC: American Council on Education.

Jaeger, R.M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10,* 3-6, 10, 14.

Kane, M. *(1993).* The validity of performance standards. Unpublished manuscript.

National Academy of Education (1992). *Assessing student achievement in the states. The first report of the National Academy of Education panel on evaluation of the NAEP trial state assessment: 1990 trial state assessment.* Stanford, CA: Author.

National Assessment of Educational Progress Authorization Act, (P.L. 110-279).

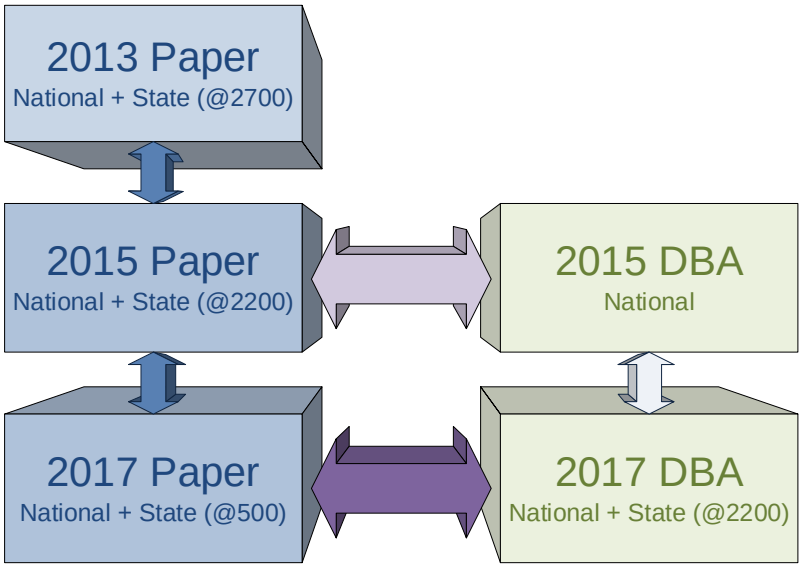
Reid, J.B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice, 10,* 11-14.

Endnotes

1. The traditional scale anchoring procedures anchored at the 200, 250, 300 350 points of the scale (± 12.5 points), using a p = *.65,* and a discrimination of .30 with the next lower level. The modified anchoring procedures (tried in reading for 1992) anchored at the achievement levels cut scores ***(±. 12.5),*** using a p = .65, and no discrimination criterion.

2017 Digital-Based Assessment (DBA) design and update on analysis timeline

In 2017, the NAEP *operational* reading and mathematics assessments (grades 4 and 8) are transitioning to a digital platform. The graphic below provides an overview of the various samples and conditions involved, where 'State' is to mean both state and urban districts. In the graphic, indicated state sample sizes are approximate and depend on urban district presence (within a state) and size.



As the graphic reveals, the design includes two bridge studies: one in 2015 and one in 2017. Data collection for the first bridge study was part of the 2015 paper-based operational administration and entailed additional national samples in all three grades for mathematics, reading, and science. In these additional national samples, a tablet-based version of the various NAEP instruments was administered on NAEP-provided tablets. The second bridge study is currently in the field for data collection and spans math and reading in 4th and 8th grade. The administration contains small state-level samples of 500 students per state participating in the paper-based assessment alongside larger state-level samples of 2200 students per state participating in the tablet-based assessment. In addition, trial urban districts are included with samples of at least 500 students each in both the paper and tablet modes. Together with private school samples in both modes, the state and urban district samples, in aggregation, yield very sizeable and randomly equivalent national samples across modes.

At the highest level, there are two chained questions to be answered:

1. Do we measure the same construct across modes?
2. If so, are (construct-irrelevant) mode differences constant across student groups?

Answering those questions is complicated and key factors that are brought to bear include dimensionality, national student group differences, state-level differences, and stability over time. The 2015 national-only study provided some insights into the data, which helped inform the 2017 study. The 2017 state-level study has been designed to provide answers about the robustness of the results over time and individual state and urban district differences.

A high level 2017 analysis **timeline** with milestones is listed below:

• **Late May, 2017**	Receive scored responses and final weight files. Start analyses for reading and mathematics at grades 4 and 8.
• **June, 2017**	Conduct initial descriptive summary statistics for reading and mathematics at grades 4 and 8; produce observed-score results comparing paper and tablet based percent correct and missing rates by various student groups, within and across years (2015 and 2017).
• **July to early September, 2017**	Generate equated scale score results for reading and mathematics at grades 4 and 8 for paper and tablet based assessments, separately; disaggregate and compare the paper and tablet based scale score results by student groups and by state/urban district, within and across years. Technical review.
• **Mid-September, 2017**	Include results in draft reports, report review cycles continue and supplementary materials may be developed to inform constituents about the mode transition in more detail.
• **Late September to December, 2017**	Conduct additional analyses to support the 2017 reporting.

Reporting Relevant Progress: Exploring Dynamic Frameworks for NAEP

Overview

According to the NAEP statute (P.L. 107-279), the Governing Board is responsible for developing assessment objectives and test specifications for each NAEP subject area. Since 1989 the Governing Board has developed assessment frameworks and specifications in more than 10 subjects through comprehensive, inclusive, and deliberative framework projects. The Board's Framework Development Policy is included in this attachment.

Three models have been used to account for the need to update framework content over time:

1. New Framework/Start New Trend

In some cases, the Board has determined through research, outreach, content and policy input, and other means that a new framework is warranted in a subject area. In these subject area assessments, the new assessment framework defines a new construct, includes different content and skills, adds new item types, changes the assessment delivery mode (i.e., digital-based assessment (DBA)), and other modifications. Examples of this model include 2011 NAEP Writing, where the new construct was writing on a computer and using word processing tools. This was judged to represent a different construct from writing in the previous framework's paper and pencil assessment. The new construct definition motivated a break in trend reporting from the old assessment's results. A similar break in trend occurred for the 2009 NAEP Science Framework, which reflected several enhancements from advancements in science and science curricular standards, such as crosscutting content and deeper integration of science practices.

2. New Framework/Maintain Trend

In this model, the new framework is designed to be different in many ways from the previous framework; however, empirical investigation reveals that the construct does not differ substantially. The interest in maintaining trend prompts linking studies and other research to try to ensure trend lines can be maintained. Board adoption of the 2009 NAEP Reading Framework was under similar circumstances as the new frameworks for NAEP Writing and NAEP Science, because the old NAEP Reading Framework had several sub-elements that were no longer relevant to the field's conceptualization of reading comprehension. This framework update occurred during the No Child Left Behind Act (NCLB) era. Given the NCLB statute's requirement to use NAEP as a monitoring tool for states, there was substantial interest in establishing a bridge to maintain the reading trend despite changes to the construct being measured on NAEP. Empirical investigation revealed that trend reporting could be maintained, and so the NAEP Reading trend remained intact from its beginning in 1992.

3. Updated Framework/Maintain Trend

This model is defined by gradual changes to a framework over time so that trend is maintained. For mathematics, the framework has been "tweaked" over time to more clearly define the objectives, shift content emphases, and refine the process dimension while not redefining the construct. NAEP has been able to maintain the mathematics trend line for grades 4 and 8 since 1990. The framework "tweaks" have occurred sporadically rather than on an ongoing basis, often prompted by less dramatic but important curricular and assessment advances for a subject area. A more ongoing and systematic model for these updates could be included in the concept of dynamic frameworks.

Dynamic Framework Model

The Board's Strategic Vision, adopted at the November 2016 quarterly meeting, includes a goal to:

> Develop new approaches to update NAEP subject area frameworks to support the Board's responsibility to measure evolving expectations for students, while maintaining rigorous methods that support reporting student achievement trends.

This description in the Strategic Vision suggests a fourth model for making continuous, gradual changes to NAEP frameworks using empirical evidence to avoid compromising the ability to maintain trend. This more systematic and ongoing approach to updating assessment content is novel and has been referred to as *dynamic frameworks*. First described in The Future of NAEP (attached), a dynamic framework incorporates continuous, incremental changes to content rather than periodic abrupt shifts in content.

According to The Future of NAEP:

> "Dynamic frameworks would balance dual priorities of trend integrity and trend relevance. As an analogy, the Consumer Price Index (CPI) tracks inflation by deliberately conflating two concepts: change in the cost of a fixed basket of goods and change in the composition of the basket itself. As time passes, an increase in the cost of a product that is no longer relevant should contribute less to estimated inflation. By adopting dynamic frameworks, NAEP would similarly conflate increases in student proficiency with a change in the definition of proficiency itself. Although this conflation may seem undesirable, it may be the best way to balance desires for both an interpretable trend and a relevant trend" (p. 17).

There are several issues and questions that need to be resolved before the Governing Board can make a determination about the feasibility of a dynamic framework model. Issues related to the reasons for updating frameworks and what content to add or delete is in the domain of the Assessment Development Committee (ADC). Issues related to the speed of change and methods for maintaining trend with continuous, incremental changes to content would be in the domain of the Committee on Standards, Design and Methodology (COSDAM) and the National Center for Education Statistics (NCES).

During the November 2016 quarterly Board meeting, the ADC and COSDAM met to begin discussing how to approach the idea of dynamic frameworks. An excerpt of the minutes from that joint committee meeting discussion is included in this attachment. The committees agreed that additional time was needed to discuss how to approach a dynamic framework model.

The following suggested discussion questions are provided to support the March 2017 joint session of the committees and address some of the issues involved in pursuing this approach.

Discussion questions

1. What are the conceptual differences between the dynamic framework and the other 3 models (New Framework/New Trend; New Framework/Maintain Trend; Updated Framework/Maintain Trend) that the Board and NCES have employed in the past?

   Some possible factors to consider:
   - The extent to which proposed content changes are judged to represent a new construct
   - An a priori decision about the importance of maintaining trend, versus deciding whether trend can be maintained post hoc
   - The scope of proposed content changes
   - The speed of proposed content changes
   - The extent to which the approach distinguishes between adding versus dropping objectives
   - The extent to which the operationalization of the frameworks (i.e., item specifications and item pools) changes over time

2. What does an assessment development schedule currently look like?
   - The Board, through ADC, currently decides on the scope of proposed content changes through contractors that convene educators, parents, and the general public, for active and broad participation. The current framework development policy states that frameworks and test specifications shall remain stable for at least 10 years.
     - Can we make the scope of changes and related outreach more continuous, smooth, and systematic?
   - Under current operational procedures, it takes approximately 4.5 years between Board adoption of a framework (or changes to a framework) and NCES development and administration of new items under that framework.
     - Can we make this transition more continuous, smooth, and systematic?
   - With each operational administration of an assessment, several items are released and replacement items are developed by NCES. The released and replaced items may vary somewhat in terms of the objectives covered.
     - Are there implications for this process under a dynamic framework model?

3. What are the "must haves" for dynamic frameworks?
   - For example, should we posit that we must document and communicate any changes to the public and various stakeholders clearly and by a certain time?
   - For example, should we commit to upholding the current framework development policy to include the active participation of educators, parents, and members of the general public?

4. Would the possible updates being considered for the NAEP Mathematics Framework be a good time to try out a dynamic framework model?

   Possible next steps related to assessment content:

   - Determine whether there is a compelling rationale to pursue content updates
   - Determine which objectives (if any) should be dropped (procurement underway)

- Determine which objectives (if any) should be added (procurement underway)
- Identify how to ensure content updates are determined through an inclusive and deliberative process with active participation from states, educators, and parents
- Determine how quickly a revised framework can be developed and adopted
- Gather information about the number of items related to dropped objectives that have appeared on recent NAEP assessments
- Determine how quickly an assessment can be administered based on a revised framework that drops some objectives
- Determine how quickly an assessment can be administered based on a revised framework that adds some new objectives
- Determine whether each increment of change is meaningful and defensible from a content perspective

Possible next steps related to methodology:

- Reassess Board commitment to maintaining trends in NAEP Mathematics in 2021 and beyond
- Determine how (if at all) recent NAEP results would have been different if the assessment had not included items associated with objectives that will be dropped
- Determine what factors affect the speed with which a framework can be revised while still maintaining trends:
    i. Number or proportion of objectives to be deleted
    ii. Number or proportion of objectives to be added
    iii. Number or proportion of items associated with deleted/added objectives
    iv. Difficulty of items associated with deleted/added objectives
    v. Potential changes to other aspects of the framework (e.g., cognitive processes)
    vi. Other
- Consider implications of content changes for achievement levels

Except of minutes from November 2016 joint session on Dynamic Frameworks

Mary Crovo provided an overview with historical context about ways in which the Board has changed frameworks while maintaining or breaking trend lines. In these instances, NAEP has either continued to report trends on new assessment results connecting with previous results or started a new reporting trend relative to previous assessment results. She noted that NAEP's practice has been to reflect broad-based input from many stakeholders. Ms. Crovo summarized there are three different ways that NAEP has dealt with framework changes: starting a new framework and breaking the trend line for the assessment results; starting a new framework and maintaining the trend line connecting to the previous framework; and implementing smaller framework updates while maintaining the trend line.

Ms. Crovo also reviewed the current timeline for development of an assessment, from framework development to reporting of results. Joe Willhoft made a note of the long lead time of nearly 4.5 years between a framework's completion and the final operational assessment being administered, but Ms. Crovo noted that smaller or more incremental framework changes could shorten this timeline with fewer items to develop.

As part of this session, the Committees also heard a presentation from Dan McGrath of NCES to summarize how NCES has considered the concept of dynamic frameworks for NAEP as part of the NCES Future of NAEP initiative, and how international assessments have approached this concept of updating frameworks.

Cary Sneider noted that the Board could foreseeably identify rationales for shifting the percentages of content or having content that repeats in multiple grades. For example, such changes could address cases where there are NAEP alignment issues resulting primarily from different sequencing of content across grades, and these changes provide helpful information on how learning progresses on the same content, from grade 4 to 8.

Lucille Davy noted that the grade 4 NAEP Mathematics Assessment has some content most students are not learning by the 4th grade, as indicated by several states' adoption of Common Core State Standards. She acknowledged the need to study how much change is too much and to study the ideal rate of change over time, in order to optimize both measurement of student performance and relevance to education policy.

Dale Nowlin commented that even when we do not change the measure, i.e., the assessment, what is being measured is changing. The NAEP Writing Assessment shows this clearly—the current NAEP Writing Framework reflects a construct focused on writing in a digital environment with common word processing tools, but if NAEP continued to assess students in the traditional paper-pencil format today, the assessment would not collect the same information compared to the student performance data gathered from the last paper-pencil assessment because this is increasingly not the way students write.

In addition to the rate of implemented changes, the Committees noted several issues that need to be carefully considered and balanced. Mitchell Chester suggested reviewing how shifting the context of items can represent desired changes, without changing the construct. Ms. Garrison noted that time limitations for assessment administrations are an important factor, as well as assuring that current NAEP items remain relevant to students in future administrations. Joe Willhoft suggested we examine how new changes may interact with general content drift over

time or the accumulation of year-to-year trend inferences over time. Finally, Linda Rosen and Mr. Willhoft noted that different stakeholders may react to changes differently.

Mr. Willhoft also noted that the Board should carefully consider how communications with educators are framed so that messages do not create a sense that students are chasing a moving target, with an assessment that is constantly changing. Jim Popham encouraged the Board to promote educational progress in how the concept of dynamic NAEP assessment frameworks is defined and pursued.

Several Committee members agreed on the importance of clarifying and articulating the problem that the Board is hoping to address with a dynamic assessment framework model. Mr. Chester asked the Board to consider changes in the field that NAEP is not detecting in the current more static framework model, and whether these changes are important for NAEP to capture. Generally, the Committees agreed about the need to study how much change is too much, i.e., what level of change would potentially compromise NAEP's ability to report trends over time. Another important issue is how to implement proposed changes.

The framework updates that the Board will eventually consider for the NAEP Mathematics Assessment will be a first case where the concept of dynamic frameworks can be applied. Ms. Crovo noted that the Board is commissioning research to comprehensively survey state mathematics standards, including the 15 percent of additional state-level standards. This research will inform decisions on whether and how to change the current NAEP Mathematics Framework.

Ms. Davy also reminded the Committee that several of these issues are time sensitive to best support states, and so Board discussion should be deliberate and also reflect this urgency. Chasidy White agreed that states need guidance on these issues. The Committees requested continued joint Committee discussion to grapple with these issues and open questions, with a next meeting that focuses more on understanding current processes and considering how they could be changed.

Adopted: May 18, 2002

National Assessment Governing Board

Framework Development

Policy Statement

It is the policy of the National Assessment Governing Board to conduct a comprehensive, inclusive, and deliberative process to determine the content and format of all subject area assessments under the National Assessment of Educational Progress (NAEP). Objectives developed and adopted by the Governing Board as a result of this process shall be used to produce NAEP assessments that are valid and reliable, and that are based on widely accepted professional standards. The process shall include the active participation of educators, parents, and members of the general public. The primary result of this process shall be an assessment framework to guide NAEP development at grades 4, 8, and 12.

The Governing Board, through its Assessment Development Committee, shall carefully monitor the framework development process to ensure that all Governing Board policies are followed; that the process is comprehensive, inclusive, and deliberative; and that the final Governing Board-adopted framework, specifications, and background variables documents are congruent with the Guiding Principles, Policies, and Procedures that follow.

Introduction

Since its creation by Congress in 1988, the Governing Board has been responsible for determining the content and format of all NAEP subject area assessments. The Governing Board has carried out this important statutory responsibility by engaging a broad spectrum of educators, policymakers, business representatives, and members of the general public in developing recommendations for the knowledge and skills NAEP should assess in various grades and subject areas. From this comprehensive process, the Governing Board develops an assessment framework to outline the content and format for each NAEP subject area assessment.

Under provisions of the National Assessment of Educational Progress Authorization Act of 2002 (P.L. 107-279), Congress has authorized the Governing Board to continue its mandate for determining the content and format of NAEP assessments by requiring that:

- "the purpose [of NAEP] is to provide…a fair and accurate measurement of student academic achievement;"
- "[NAEP shall]…use widely accepted professional testing standards, objectively measure academic achievement, knowledge, and skills, and ensure that any academic assessment authorized….be tests that do not evaluate or assess personal or family beliefs and attitudes or publicly disclose personally identifiable information;"
- "[NAEP shall]…only collect information that is directly related to the appraisal of academic achievement, and to the fair and accurate presentation of such information;"
- "the Governing Board shall develop assessment objectives consistent with the requirements of this section and test specifications that produce an assessment that is valid and reliable, and are based on relevant widely accepted professional standards;"
- "the Governing Board shall have final authority on the appropriateness of all assessment items;"
- "the Governing Board shall take steps to ensure that all items selected for use in the NAEP are free from racial, cultural, gender, or regional bias and are secular, neutral, and non-ideological;" and
- "the Governing Board shall develop a process for review of the assessment which includes the active participation of teachers, curriculum specialists, local school administrators, parents, and concerned members of the public."

Given the importance of these mandates it is incumbent upon the Governing Board, in the design, conduct, and final action on the assessment framework, to ensure that the highest standards of test development are employed. The validity of educational inferences made using NAEP data could be seriously impaired without high standards and rigorous procedures for framework development.

Historically, the task of developing the framework for a NAEP assessment has been conducted by the Governing Board through competitive procurements. It is imperative that contractors be fully informed of the Governing Board's policy regarding framework development, so that all deliverables under the contract meet statutory requirements and are acceptable to the Governing Board. The purpose of the Policy on Framework Development, therefore, is to articulate the Guiding Principles, Policies, and Procedures that will direct the framework development process.

Each of the following Guiding Principles is accompanied by Policies and Procedures. Full implementation of this framework development policy will require the appropriate framework contractor(s), to provide assurances to the Governing Board, through the Governing Board staff, that all aspects of the Policies and Procedures for which they are responsible have been successfully completed. These assurances will be in writing, and may require supporting information prepared by the contractor and/or the Governing Board staff.

This policy complies with the documents listed below which express widely accepted technical and professional standards for test development. These standards reflect the agreement of recognized experts in the field, as well as the policy positions of major professional and technical associations concerned with educational testing.

Standards for Educational and Psychological Testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

Code of Fair Testing Practices in Education. (2004). Washington, DC: Joint Committee on Testing Practices.

National Center for Education Statistics (NCES) Statistical Standards, September 2002.

Guiding Principles – Framework Development

Principle 1

The Governing Board is responsible for developing an assessment framework for each NAEP subject area. The framework shall define the scope of the domain to be measured by delineating the knowledge and skills to be tested at each grade, the format of the NAEP assessment, and preliminary achievement level descriptions.

Principle 2

The Governing Board shall develop an assessment framework through a comprehensive, inclusive, and deliberative process that involves the active participation of teachers, curriculum specialists, local school administrators, parents, and members of the public.

Principle 3

The framework development process shall take into account state and local curricula and assessments, widely accepted professional standards, exemplary research, international standards and assessments, and other pertinent factors and information.

Principle 4

The Governing Board, through its Assessment Development Committee, shall closely monitor all steps in the framework development process. The result of this process shall be recommendations for Governing Board action in the form of three key documents: the assessment framework; assessment and item specifications; and background variables that relate to the subject being assessed.

Principle 5

Through the framework development process, preliminary achievement level descriptions shall be created for each grade being tested. These preliminary descriptions shall be an important consideration in the item development process and will be used to begin the achievement level setting process.

Principle 6

The specifications document shall be developed during the framework process for use by NCES and the test development contractor as the blueprint for constructing the NAEP assessment and items in a given subject area.

Principle 7

NAEP assessment frameworks and test specifications generally shall remain stable for at least 10 years.

Policies and Procedures for Guiding Principles

Principle 1

The Governing Board is responsible for developing an assessment framework for each NAEP subject area. The framework shall define the scope of the domain to be measured by delineating the knowledge and skills to be tested at each grade, the format of the NAEP assessment, and preliminary achievement level descriptions.

Policies and Procedures

1. The assessment framework shall determine the extent of the domain and the scope of the construct to be measured for each grade level in a NAEP assessment. The framework shall cover grades 4, 8, and 12, where applicable, in a given subject area. The framework shall provide information to the public and test developers on three key aspects of the assessment: a) what should be measured; b) how that domain of content is most appropriately measured in a large-scale assessment; and c) how much of the content domain, in terms of knowledge and skills, should students know and be able to do at the basic, proficient, and advanced levels.

2. More specifically, the framework shall: a) articulate the purpose and scope of the assessment; b) define the content and skills to be tested at each grade; c) define the weighting of the item pool in terms of the content and process dimensions; d) describe the format requirements of the items and the assessment; e) include preliminary achievement level descriptions for each grade at the basic, proficient, and advanced levels; and f) contain sample items for each grade to be tested.

3. The primary audience for the assessment framework shall be the general public. Technical and subject-specific terminology should be used only when necessary, and shall be defined in the body of the framework or in a glossary. Where appropriate, the framework should use tables, charts, and graphics to clearly and concisely communicate necessary information pertaining to the various assessment elements. The framework shall contain sufficient information to inform policymakers, educators, and others about the nature and scope of the assessment in a given subject area.

4. NAEP frameworks shall continue to be developed with the active participation of states. Content coverage in each subject and grade shall be broad, inclusive of content valued by states as important to measure, and reflect high aspirations for student achievement.

5. The framework shall not endorse or advocate a particular pedagogical approach to the subject area being assessed, but shall focus on important, measurable indicators of student achievement to inform the nation about what students know and are able to do. While the framework shall not endorse pedagogy, it may facilitate reporting on various types of skills essential to achievement in the grade and subject area.

6. Where appropriate, the framework shall describe additional requirements of the assessment and administrative conditions which may be unique to a given subject area. For example, this may include a brief discussion of ancillary materials, use of technology, and other conditions.

7. Special studies, if any, to be conducted as part of the assessment in a given subject area shall be described in the framework. This description shall provide an overview of the purpose and rationale for the study, the nature of the student sample(s), and a discussion of the instrument and administration procedures.

8. Following Governing Board adoption, the framework shall be widely disseminated in print and electronic versions.

Principle 2

The Governing Board shall develop an assessment framework through a comprehensive, inclusive, and deliberative process that involves the active participation of teachers, curriculum specialists, local school administrators, parents, and concerned members of the public.

Policies and Procedures

1. The guiding statute calls for the "active participation" of various NAEP audiences in the framework development process. Because this is a public endeavor it is important that all major constituents are represented in a fair and open process. The Governing Board's framework development process shall be comprehensive in its scope and outreach; inclusive in its involvement of broad-based panel members and reviewers; and deliberative in considering all viewpoints and debating all pertinent issues in formulating the content and design of a NAEP assessment.

2. The framework development committees shall be constituted in such a way as to be representative in terms of gender, race/ethnicity, region of the country, and viewpoints regarding the content of the assessment under development. In addition, many different views shall be sought from various segments of the population in the review of materials and in soliciting public input and feedback. The level of "active participation" shall be documented in a report of the framework development process.

3. The framework development environment shall be open, balanced, and even-handed. To the greatest extent possible, the project deliberations will be protected from inappropriate influences of various interest groups. All issues and agendas shall be considered in a careful, objective, and respectful manner by all project committees and the Governing Board.

4. Prior to implementation of the framework development process, the contractor shall identify procedures that will be used to clarify positions and views, roles and responsibilities of all project staff and committees, as well as how the process will work toward reaching an understanding of the scope, content, and design of the framework.

5. While the NAEP statute no longer requires a "national consensus process," the Governing Board will develop frameworks through involvement of broadly representative groups and individuals with diverse viewpoints, open discussion and deliberation of issues, and careful consideration, and revision when necessary, of framework recommendations prior to final Governing Board action. The Governing Board shall make the final decision on a framework and shall not delegate decisions on the content and format of NAEP assessments.

6. It is a requirement throughout the framework development process to obtain reviews of draft materials and general public input from a wide audience of stakeholders, including content experts (outside of the framework committees), curriculum and assessment staff of state and local education agencies, users of assessment data, those who are employed in the specific content area under consideration, policymakers, parents, and the general public. The constituency of "users and consumers" mentioned above may include scientists, mathematicians, journalists, civic leaders, authors, and others.

7. Written summaries of all hearings, forums, surveys, and committee meetings shall be made available to the framework committees in a timely manner, so that such information can best inform the decisionmaking process. The Assessment Development Committee and the Governing Board shall receive written documentation and regular briefings on all project activities at their quarterly meetings.

8. Framework development panels shall consist of a policy oversight or steering committee comprised of representatives from key policy groups, business and industry, content experts, educators at the state and district level, users and consumers, parents, and the general public. At least 30 percent of this committee shall be composed of users and consumers in the subject area under consideration. Both public and private schools shall be represented on this committee.

9. The steering committee will receive the project charge directly from the Governing Board, and shall formulate guidelines for the conduct of the framework development process, consistent with statutory requirements and Governing Board policy. This oversight committee shall monitor the progress of the development work via meetings, teleconferences, and electronic communication. The final recommended documents from the project shall be reviewed by the oversight panel for recommendation to the Governing Board at the completion of the deliberative process.

10. Development of the project documents shall be the responsibility of a project planning committee composed of content experts, educators at the state and district level, curriculum specialists, university professors, policymakers, users and consumers, business representatives, and members of the public. Classroom teachers shall be well represented on this committee at all grade levels designated for the assessment under development. Teachers, administrators, and curriculum specialists shall be drawn from schools across the nation, including individuals who work with students from high-

poverty and low-performing schools. Both public and private schools shall be represented on this committee.

11. The planning committee shall carefully consider the charge from the Governing Board and guidelines set forth by the project oversight committee in developing the assessment framework. The committee shall carry out its work through meetings, conference calls, and electronic communication. It shall be responsible for developing the major deliverables of the project: the framework, specifications, and background variables documents, under the direction of project staff.

12. Where appropriate, a third committee of technical experts shall be involved in the framework development process. This committee shall consist of psychometricians, state testing experts, and individuals involved in developing assessments in the content area under consideration. It shall be this panel's responsibility to uphold the highest technical standards for development of the NAEP framework and specifications. The committee shall respond to technical issues raised during the process and provide guidance to project staff and the project committees on technical aspects of the assessment specifications. As with the steering and planning committees, the technical panel will meet in-person, via teleconference, and through electronic communication.

13. The preceding Policies and Procedures for conducting the framework development process constitute one model of committee structure. A prospective contractor may propose an alternative plan; however, the committees must be broad-based and representative of the type of groups and individuals identified above.

Principle 3

The framework development process shall take into account state and local curricula and assessments, widely accepted professional standards, exemplary research, international standards and assessments, and other pertinent factors and information.

Policies and Procedures

1. The NAEP framework development process shall be informed by a broad, balanced, and inclusive set of factors. The framework shall maintain a balance between curriculum reform in a field, exemplary research regarding cognitive development and instruction, and the nation's future needs and desirable levels of achievement. This delicate balance between "what is" and "what should be" is the essence of the NAEP framework development process.

2. The framework development process shall begin by thoroughly identifying major policy and assessment, issues in the content area, to be summarized in an issues paper. The primary audiences for the issues paper are the Governing Board and the project committees. Designed to serve as a springboard for committee deliberations and framework development, this paper shall elaborate on major issues providing both pros and cons, summarize the research, and cite trends in state standards and assessments.

3. The framework panels shall consider a wide variety of resources as the deliberations proceed, including but not limited to curriculum guides and assessments developed by states and local districts, widely accepted professional standards, scientific research, other types of exemplary research studies in the literature, key reports having significant national and international interest, international standards and assessments, other assessment instruments in the content area, and prior NAEP frameworks, if available.

4. In considering the relative importance of these sources of information in developing the framework, the project committees shall consider the charge as delivered by the Governing Board, the role and purpose of NAEP in informing the public about student achievement, constraints of a large-scale assessment, technical assessment standards, issues of burden and cost-effectiveness in designing the assessment, and other factors unique to the content area.

Principle 4

The Governing Board, through its Assessment Development Committee, shall closely monitor all steps in the framework development process. The result of this process shall be recommendations for Governing Board action in the form of three key documents: the assessment framework; assessment and item specifications; and background variables that relate to the subject being assessed.

Policies and Procedures

1. When the framework development process is conducted for the Governing Board by an outside contractor, the process shall be managed in an efficient, cost-effective manner, shall be completed in a timely fashion, and shall adhere to sound measurement practice.

2. The Governing Board's Assessment Development Committee (ADC) shall be responsible for monitoring the framework development process that results in recommendations to the Governing Board on the content and format of each NAEP assessment. Direction will be provided to the framework development contractor by the ADC and the Governing Board, via Governing Board staff, to assure compliance with the NAEP law, Governing Board policies, Department of Education and government-wide regulations, and requirements of the framework contract.

3. The performance of work for the framework development process shall be subject to the technical direction of a Governing Board staff member, designated as the Contracting Officer's Representative. This individual shall work under the guidance of the ADC and the Governing Board during all phases of the framework process.

4. During the framework process, the Governing Board shall review work-in-progress and make modifications as necessary. The Governing Board shall receive regular updates on the framework development process at its quarterly meetings. Updates

shall be provided to the ADC as necessary during the framework development process via in-person meetings, teleconferences, printed material, and electronic communication.

5. At the conclusion of the framework development process, the Governing Board will take final action on the recommended framework, specifications, and background variables documents. This action may result in modifications to one or more of the documents, which will be incorporated prior to dissemination.

6. The framework process shall also result in recommendations to the Governing Board on background variables to be collected from students, teachers, and schools related to a particular subject area. Such variables shall be related to academic achievement and to the fair and accurate presentation of achievement information. Background variables shall meet criteria for being secular, neutral, and non-ideological, as stated in the Governing Board's Policy on NAEP Item Development and Review, and will not assess personal or family beliefs and attitudes, or publicly disclose personally identifiable information. In recommending background variables, the Governing Board's Policy on Collecting and Reporting Background Data shall also be followed. Recommendations on background variables shall take into account burden, cost, quality of the data to be obtained, and other factors.

7. Following adoption by the Governing Board, the final framework, specifications, and background variables documents shall be provided to NCES at least 12 months prior to pilot or field testing, except in the case of unforeseen circumstances related to congressional action, budget limitations, or other extraordinary events.

Principle 5

Through the framework development process, preliminary achievement level descriptions shall be created for each grade being tested. These preliminary descriptions shall be an important consideration in the item development process and will be used to begin the achievement level setting process.

Policies and Procedures

1. The framework panels shall draft preliminary descriptions for basic, proficient, and advanced performance for all applicable grades in the content area under development. The panels shall use the Governing Board's policy definitions for basic, proficient, and advanced achievement in developing the preliminary descriptions. The descriptions shall provide statements of what students should know and be able to do, as derived from the content and process dimensions of the assessment at each grade.

2. The preliminary descriptions shall be included in the framework draft that is widely circulated for public review and comment, to obtain broad input on the draft descriptions prior to Governing Board action on the framework.

3. Once the Governing Board has approved the framework document, NCES shall be provided with the preliminary achievement levels descriptions so that these definitions can guide development of NAEP test questions.

4. The preliminary descriptions approved by the Governing Board shall also be provided to the achievement levels contractor to begin the level-setting process.

Principle 6

The specifications document shall be developed during the framework process for use by NCES and the test development contractor as the blueprint for constructing the NAEP assessment and items in a given subject area.

Policies and Procedures

1. The assessment and item specifications shall produce an assessment that is valid and reliable, and based on relevant widely accepted professional standards. The specifications shall also be consistent with Governing Board policies regarding NAEP design such as booklet and block (item sets within a booklet) structure, test administration conditions, and accommodations for special needs students.

2. The primary audience for the specifications, or assessment blueprint, shall be the contractor(s) responsible for developing the assessment and test questions. The specifications shall be written in sufficient detail so that item writers can develop high-quality questions based on the framework objectives for grades 4, 8, and 12, where applicable, in a given subject area.

3. The specifications shall include, but not be limited to: a) detailed descriptions of the content and process dimensions, including the weighting of those dimensions in the pool of questions at each grade; b) types of items; c) guidelines for stimulus material; d) types of response formats; e) scoring procedures; f) preliminary achievement level descriptions; g) administration conditions; h) description of ancillary or additional materials, if any; i) considerations for special populations; j) detailed information on special studies, if any; k) a substantial number and range of sample items with scoring guidelines for each grade level; and l) any unique requirements for the given subject area.

4. The specifications shall evolve from the framework document, and be carefully reviewed by technical experts involved in the process, prior to submission to the Governing Board.

Principle 7

NAEP assessment frameworks and test specifications generally shall remain stable for at least 10 years.

Policies and Procedures

1. Development of a new subject area framework shall be guided by the schedule of NAEP assessments adopted by the Governing Board.

2. In deciding when to conduct a new framework development process for an existing NAEP assessment, the Board shall consider factors such as exemplary research, curriculum and assessment reform, widely accepted professional standards, implications for existing trendlines, cost and technical issues, and other factors.

3. In rare circumstances, such as where significant changes in curricula have occurred, the Governing Board may make changes to assessment frameworks and specifications before 10 years have elapsed.

4. In those subjects and grades for which NAEP would provide confirmatory evidence about progress in achievement on state tests, the Governing Board shall revise frameworks only when the rationale for doing so is compelling.

MAY 2012

NAEP:
LOOKING AHEAD

LEADING ASSESSMENT INTO THE FUTURE

Recommendations to the Commissioner
National Center for Education Statistics

NAEP: Looking Ahead
Leading Assessment into the Future

NCES Initiative on the Future of NAEP

The National Assessment of Educational Progress (NAEP) has undergone a series of notable changes in the past decade. The NAEP program has expanded to meet new demands. All 50 states, the District of Columbia, the Department of Defense schools, and (on a trial basis) 21 urban districts are now participating in the mathematics and reading assessments at grades 4 and 8. In addition, thirteen states are participating in trial state 12th-grade assessments in reading and mathematics. NAEP is also reporting in record time to ensure that the findings are highly relevant upon release. Technology has taken on a bigger role in the development and administration of NAEP, including computer-based tasks in the science and writing assessments. These are just a few of the major developments; the program has grown and matured in almost all respects.

There is also growing interest in linking NAEP to international assessments so that NAEP scores can also show how our nation's students measure up to their peers globally. Additionally, there is increasing interest in broadening assessments in the subject areas to incorporate college and career readiness, as well as what are often called "21st-century skills" (communication, collaboration, and problem-solving).

The National Center for Education Statistics (NCES), which administers NAEP, is dedicated to moving the program forward with its upcoming procurement cycle which will take the program to 2017. Under the leadership of NCES Commissioner Jack Buckley, NCES convened a diverse group of experts in assessment, measurement, and technology for a summit in August 2011. These experts discussed and debated ideas for the future of NAEP. NCES convened a second summit of state and local stakeholders in January 2012. Participants at both gatherings were encouraged to "think big" about the role that NAEP should play in the decades ahead.

NCES assembled a panel of experts from the first summit, chaired by Edward Haertel, an expert in educational assessment, to consider and further develop the ideas from the two discussions and make recommendations on the role of NAEP in the future—10 years ahead and beyond. Based on summit deliberations and their own extensive expertise, the panel developed a high-level vision for the future of the NAEP program, as well as a plan for moving toward that vision.

This paper contains the panel's recommendations to the NCES Commissioner. NCES will consider these recommendations in their mid- and long-range planning for the program.

Panel Members

Edward Haertel (chair)
Jacks Family Professor of Education
School of Education
Stanford University

Russell Beauregard
Research Scientist & Director of Design
Education Market Platforms Group
Intel Corporation

Jere Confrey
Joseph D. Moore Distinguished University Professor
College of Education
North Carolina State University

Louis Gomez
MacArthur Chair in Digital Media and Learning
Graduate School of Education and Information Sciences
University of California, Los Angeles

Brian Gong
Executive Director
National Center for the Improvement of Educational Assessment

Andrew Ho
Assistant Professor
Graduate School of Education
Harvard University

Paul Horwitz
Senior Scientist and Director
Concord Consortium Modeling Center
Concord Consortium

Brian Junker
Professor
Department of Statistics
Carnegie Mellon University

Roy Pea
David Jacks Professor of Education
School of Education
Stanford University

Robert Rothman
Senior Fellow
Alliance for Excellent Education

Lorrie Shepard
Dean and Distinguished Professor
School of Education
University of Colorado, Boulder

3. NAEP's Assessment Frameworks and Learning Outcomes

3.1 Background and History

Assessment frameworks are conceptual, overview documents that lay out the basic structure and content of a domain of knowledge and thereby serve as a blueprint for assessment development. Typically, assessment frameworks, for NAEP and for other large-scale assessments, are constructed as two-dimensional matrices of content strands and cognitive processes. For example, the current NAEP mathematics framework includes five content areas: number properties and operations; measurement; geometry; algebra; and data analysis, statistics and probability. These are assessed at different levels of cognitive complexity, which include mathematical abilities such as conceptual understanding, procedural knowledge, and problem-solving. In geography, the content areas include: space and Earth places; environment and society; and spatial dynamics and connections. The levels of the cognitive dimension consist of knowing, understanding, and applying.

NAEP Assessment Frameworks are developed under the auspices of the Governing Board through an extensive process involving subject matter experts, who consider how research in the discipline and curricular reforms may have shifted the conceptualization of proficiency in a given knowledge domain. The development process also requires multiple rounds of reviews by educators, policy leaders, members of the public, and scholars. It is expected that assessment frameworks will need to be changed over time. However, the decision to develop new frameworks is approached with great caution because measuring change requires holding the instrument constant. Introducing new frameworks—while providing a more valid basis for the assessment—could threaten one core purpose of NAEP, which is to monitor "progress." In the past, when relatively minor changes have been made in assessment frameworks, as judged by content experts, trend comparisons over time have been continued and bridge validity studies have been conducted to verify that conclusions about gains have not been conflated with changes in the measuring instrument or redefinition of the construct being assessed.

When more profound changes occur in the conceptualization of an achievement domain, then a new framework is essential, and correspondingly the beginning of a new trend line. The adoption by nearly all states of the CCSS in English language arts and literacy and mathematics and the new Science Education Framework developed by the National Research Council (NRC) could be the occasion for a substantial enough change in conceptualization of these domains that new NAEP frameworks and new trend comparisons are warranted. Still, the future of NAEP—as a statistical indicator and as an exemplar of leading-edge assessment technology—requires great care and attention to the implications of new trend comparisons rather than merely acceding to the hoopla surrounding the new standards.

In the history of NAEP, few changes have been made in the assessment frameworks for reading and for mathematics. The old frameworks in these two core subjects, begun in 1971 and 1973 respectively, were replaced in the early 1990s, and then again in 2009 for reading. The old assessments have been continued on a less frequent cycle and are referred to as long-term trend NAEP. The 1990's mathematics framework and 2009 reading framework guide the present-day assessments, referred to as main NAEP. While NCES has been careful to insist that the old and new frameworks measure different things and therefore cannot be compared, the existence of the two trends provides a critically important example to illustrate how changing the measure can change interpretations about educational progress (e.g., see Beaton & Chromy, 2010). The earlier assessments focused much more on basic skills. Reading passages were generally shorter compared to today's NAEP and did not require students to demonstrate so wide a range of reading skills or answer extended-response questions. In mathematics, long-term trend NAEP had a greater proportion of computational questions and items asking for recall of definitions, and no problems where students had to show or explain their work. In a 2003 study, researcher Tom Loveless complained that the new NAEP mathematics assessment exaggerated progress in mathematics during the 1990s because gains on the basic skills test over the same period were much

smaller (when compared in standard deviation units of the respective tests). Because the two assessments are administered entirely separately, Loveless then had to rely on comparisons based on the less than satisfactory item-percent-correct metric to try to track progress in subdomains of the test. A more recent study using more sophisticated methods has largely confirmed his general conclusions, but that same study has highlighted the technical challenges of comparing trends for two assessments administered under such different conditions (Beaton & Chromy, 2010).

3.2 New Approaches for Assessment Frameworks

3.2.1 Designing frameworks and assessments to evaluate directly the effects of changing domain definitions

NAEP cannot be a research program and in particular cannot be structured to investigate the effectiveness of various instructional interventions. However, it can and should be attentive to the ways that shifting definitions of subject matter competence can affect claims about progress or lack of progress (cf. Section 3.2.3). In the CCSS context, it will be especially important to pay attention directly to potential differences between consortium-based conclusions and NAEP trends. Taking this on as a role for NAEP continues its important function as a kind of monitoring instrument. For example, when some state assessment results have shown remarkable achievement gains and closing of achievement gaps, achievement trends for the same states on NAEP have helped to identify inflated claims. These disparities might exist because of teaching-the-test practices on state tests (Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998), state content or achievement standards that do not rise to NAEP levels (Bandeira de Mello, Blankenship, & McLaughlin, 2009), exclusion of low-performing students on NAEP, or lower motivation on NAEP. More direct linking by carefully accounting for the consortium frameworks within new NAEP frameworks, would allow NAEP to act somewhat like an external monitor for CCSS assessment results. While the current NAEP frameworks do cover many of the same skills as the CCSS, they can be enhanced with some shifts in content.

"21st-century skills" aren't actually new in this century, but it is a relatively new idea (beginning in the 1990s) that these reasoning skills should be more broadly attained and expected of all students. More importantly, it is indeed new that policy leaders would move toward a view of learning that calls for reasoning and explaining one's thinking from the earliest grades, in contrast to outmoded theories of learning predominant in the 20th century that postponed thinking until after the "basics" had been mastered by rote. In addition, the CCSS firmly ground reasoning, problem-solving, and modeling in relation to specific content, not as nebulous generalized abilities. While there is widespread enthusiasm for designing new assessments that capture these more rigorous learning goals, we should note that promises like this have been made before. In the case of the current NAEP mathematics assessment, item developers acknowledge that the proportion of high complexity items actually surviving to the operational assessment is much smaller than is called for in the NAEP Mathematics Framework, and a validity study at both grades 4 and 8 found that the representation of high-complexity problems was seriously inadequate at grade 8, especially in the Algebra and Measurement strands (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007).

Good intentions to measure "higher order thinking skills" are often undermined for three interrelated reasons. First, test questions at higher levels of cognitive complexity are inherently more difficult to develop. Because the dimensions of the task are intended to be ill-specified, such problems are often perceived to be ambiguous. But as soon as the item developer provides clarifying parameters, the challenge of the problem is diminished. Second, because "21st-century skills" involve applying one's knowledge in real world contexts, prior experience with particular contexts (or lack thereof) can create very large differences in performance simply because students unfamiliar with the context are unable to demonstrate the intended content and reasoning skills. In fact, application or generalization can only be defined in relation to what is known to have been taught. This is the curriculum problem that haunts large-scale assessments like NAEP that seek to be curriculum independent. Finally, well

designed items can fail on statistical criteria if too few students can do them.

These are all cautionary tales. They do not imply that NAEP should be less ambitious in developing new assessment frameworks that reach as far as possible in representing these higher levels of subject matter proficiency. But they do suggest a hedging-one's-bets approach that does not discard old frameworks wholesale in favor of the new. Rather, as mentioned previously, some conscious combination of old and new would create an assessment better equipped to track progress over time. Later we discuss Innovations Laboratory studies like those NAEP has used historically to explore the feasibility of new assessment strategies. However, we should emphasize that studies of innovative assessment strategies that tap complex skills should not merely be new assessment formats administered to random samples of students. Rather, in recognition of the fact that opportunities to learn particular content and skills may affect whether an assessment looks psychometrically sound, studies should be undertaken with carefully selected populations where relevant opportunities to learn can be established. This will help determine whether more advanced performance can be accurately documented to exist within the parameters of the new standards.

3.2.2 Standing subject-matter panels

To aid in this process, provide substantive oversight, and ensure meaningful interpretation of trends, we elaborate a recommendation for the future of NAEP previously made by a National Academy of Education Panel, which called for standing subject-matter committees. We recommend an expanded role whereby standing committees of subject matter specialists would review field test data, for example, and call attention to instances when after-the-fact distortions of the intended domain occur because more ambitious item types fail to meet statistical criteria. These committees would also have a role in ongoing incremental updates to content frameworks. They might include at least one member with psychometric expertise to aid in formulating technical specifications. The role of these committees is further described in Section 6.1.3.

3.2.3 Dynamic assessment frameworks and reporting scales

As just explained in Section 3.1, NAEP assessment frameworks have historically been held fixed for a period of years and then changed. It might be added that historically, NAEP item pools have been constructed according to test specifications derived from assessment frameworks. NAEP reporting scales, in turn, have reflected the resulting mix of NAEP items. Periodic small revisions to assessment frameworks have been made while maintaining trend lines; major breaks requiring new trend lines have occurred only rarely. With standing subject-matter panels, assessment frameworks for each subject-grade combination might be adjusted more frequently, defining a gradually changing mix of knowledge and skills, analogous to the Consumer Price Index (cf. Section 5.3). At the same time, item pools might be expanded somewhat, including everything in the assessment framework but also covering some additional material. Assessment frameworks would still define the intended construct underlying NAEP reporting scales, but not all items in the NAEP exercise pool would be included in the NAEP reporting scales. For example, content required to maintain long-term trend NAEP, to assure sufficient representation of the CCSS, or to improve the linkage to some other assessment could be introduced into the pool without affecting NAEP reporting scales. With somewhat broader exercise pools, alternative construct definitions could be investigated in special studies. The panel assumes that broader exercise pools, supporting modestly different construct definitions, will increase the value of NAEP by highlighting distinctions among achievement patterns under different construct definitions. Of course, there would still be one main NAEP reporting scale for each subject/grade combination. Clarity in communicating NAEP findings would remain a priority.

Different assessment frameworks may imply different definitions of the same broad subject area achievement construct (e.g., "reading" or "mathematics"), and achievement trends may differ depending on the construct definition chosen. Incremental changes in assessment frameworks and the corresponding set of items on which NAEP reporting scales were based would afford local (i.e., near-term) continuity in the meaning of those scales, but over a period of decades, constructs

might change substantially. This was seen by the panel as a potential strength, but also a potential risk. Policymakers and the public should be aware of how and when the construct NAEP defines as "reading," for example, is changed. Not every small, incremental change would need to be announced, but it would be important to establish and to enforce clear policies concerning the reporting of significant changes in assessment frameworks, so as to alert stakeholders when constructs change and to reinforce the crucially important message that not all tests with the same broad content label are measuring the same thing. As small content framework adjustments accumulate over time, standing committees, using empirical studies, would need to determine when the constructs measured have changed enough to require establishing new trend lines.

Dynamic frameworks would balance dual priorities of trend integrity and trend relevance. As an analogy, the Consumer Price Index (CPI) tracks inflation by deliberately conflating two concepts: change in the cost of a fixed basket of goods and change in the composition of the basket itself. As time passes, an increase in the cost of a product that is no longer relevant should contribute less to estimated inflation. By adopting dynamic frameworks, NAEP would similarly conflate increases in student proficiency with a change in the definition of proficiency itself. Although this conflation may seem undesirable, it may be the best way to balance desires for both an interpretable trend and a relevant trend.

3.2.4 Learning progressions as possible guides to assessment frameworks

Learning progressions or trajectories represent descriptions of how students' knowledge, skills, and beliefs about the domain evolve from naïve conceptions through gradual transformations to reach proficiency with target ideas at high levels of expertise over a period of years (Heritage, 2008). They entail the articulation of intermediate proficiency levels that students are likely to pass through, obstacles and misconceptions, and landmarks, of predictable importance as students' knowledge evolves over time. Empirical study of learning progressions highlights the key roles of instruction, use of tools, and peer interactions in supporting learning. Because the process of evolving understanding can take multiple years, learning progressions bridge formative and summative assessment.

A learning progression can provide much more information than a typical assessment framework. A learning progression ideally specifies both what is to be learned as well as how that learning can take place developmentally over time. It often integrates content and cognition. It includes not only the learning targets but also common less-than-ideal states that many students pass through. It is ordered developmentally. It provides a domain-based interpretation of development or growth that is useful to educators. The 2009 NAEP Science Framework already contains a section on learning progressions; however, learning progressions may offer guidance for the development of future NAEP assessment frameworks, especially in mathematics.

Learning progressions are closely entwined with instructional decisions regarding the sequencing of key concepts and skills. In the Netherlands, for example, the related constructions are referred to as "learning-teaching trajectories." However, few empirically supported "learning progressions" as yet exist, and developing more has proven challenging. In addition, because of NAEP's role as a curriculum-independent monitor, it may be more difficult to develop assessment frameworks that are entirely built as a collection of learning progressions. More likely some particular sequences, if proven to be valid across curricula, could be embedded within more general assessment frameworks.